DEPARTMENT OF APPLIED MATHEMATICS,
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

DRAPERS' COMPANY RESEARCH MEMOIRS.

BIOMETRIC SERIES. IV.

MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION.—XVI. ON FURTHER METHODS OF DETERMINING CORRELATION.

BY

KARL PEARSON, F.R.S.

CAMBRIDGE UNIVERSITY PRESS
C. F. CLAY, MANAGER
LONDON: FETTER LANE, E.C.
EDINBURGH: 100, PRINCES STREET
H. K. LEWIS, 136, GOWER STREET, LONDON, W.C.
WILLIAM WESLEY AND SON, 28, ESSEX STREET, LONDON, W.C.
CHICAGO: UNIVERSITY OF CHICAGO PRESS
BOMBAY, CALCUTTA AND MADRAS: MACMILLAN AND CO., LIMITED
TORONTO: J. M. DENT AND SONS, LIMITED
TOKYO: THE MARUZEN-KABUSHIKI-KAISHA

1907

Price Four Shillings

3(2)

UCL0083290

University of London, University College, W.C.

The Francis Galton Laboratory
for National Eugenics.

Presented by..

Purchased..19

3(2)

LB

GALTON LAB FOLIOS 72

DEPARTMENT OF APPLIED MATHEMATICS,
UNIVERSITY COLLEGE, UNIVERSITY OF LONDON

DRAPERS' COMPANY RESEARCH MEMOIRS.

BIOMETRIC SERIES. IV.

MATHEMATICAL CONTRIBUTIONS TO THE THEORY OF EVOLUTION.—XVI. ON FURTHER METHODS OF DETERMINING CORRELATION.

BY

KARL PEARSON, F.R.S.

CAMBRIDGE UNIVERSITY PRESS
C. F. CLAY, MANAGER
LONDON: FETTER LANE, E.C. EDINBURGH: 100, PRINCES STREET
H. K. LEWIS, 136, GOWER STREET, LONDON, W.C.
WILLIAM WESLEY AND SON, 28, ESSEX STREET, LONDON, W.C.
CHICAGO: UNIVERSITY OF CHICAGO PRESS
BOMBAY, CALCUTTA AND MADRAS: MACMILLAN AND CO., LIMITED
TORONTO: J. M. DENT AND SONS, LIMITED
TOKYO: THE MARUZEN-KABUSHIKI-KAISHA

1907

Cambridge University Press

The following publications of the Department of Applied Statistics are now issued by the Cambridge University Press, viz.:—

Galton Eugenics Laboratory Publications

- I. Lecture Series.
- II. Memoir Series.
- III. Questions of the Day and the Fray.

Drapers' Company Research Memoirs

- I. Technical Series.
- Biometric Laboratory Series.
- II. Biometric Series.
- III. Studies in National Deterioration.

Copies may be obtained from:—

The Cambridge University Press, C. F. Clay, Manager; London: Fetter Lane, E.C.; Edinburgh: 100, Princes Street, or from The University of Chicago Press, Chicago, Ill., U.S.A.

Lists of the various series post free on application.

Mathematical Contributions to the Theory of Evolution.

XVI. ON FURTHER METHODS OF DETERMINING CORRELATION.

By Karl Pearson, F.R.S.

(1) *Introductory.* The object of the present paper is to give an account of some new methods of determining correlation. It is not suggested that they can with advantage replace the old processes, even when the distribution is approximately normal; to my mind the methods of determining the correlation ratio and the correlation coefficient (η and r) based on moments and product moments stand foremost for the information they give and its weighable accuracy. At the same time there are series which are so short, or cases in which it is desirable to come rapidly to an approximate result or data which cannot be presented in a form suitable for product-moment working, where other methods are not only reasonable, but necessary. To such cases the present new methods apply. I have termed them *new* methods and I think this is legitimate. In the case of the first method, I have not seen any hint of it before. In the case of what I term grade methods, Dr Spearman has suggested that rank in a series should be the character correlated, but he has not taken this rank correlation as merely the stepping stone by which to reach the true correlation of the variables as dependent magnitudes, and further in the discussion he has given of the subject he has, I believe, given erroneous formulae and made quite incorrect statements as to the magnitude of probable errors.

One word must be said as to the use made of the normal distribution. I have used it here as on many other occasions as a means of suggesting fitting relations and simple formulae for correlation constants. This does not necessarily mean (i) that the constants reached may not have a perfectly definite meaning apart from normal distributions, or (ii) that the formulae obtained may not hold for all forms of distribution apart from normality. As an illustration of the first case I cite my mean square coefficient of contingency*. This is a perfectly general measure of the deviation from independent probability in the case of an $n \times m$ fold table, but its actual form was selected so that it would agree with the coefficient of correlation in the case of indefinitely fine grouping and normal distribution. As an illustration of my second point I take the formulae given by me for the influence of selection on variation and correlation*. These formulae were originally proved for normal distributions, but for a number of years past the proofs given in my lectures have been perfectly general, depending only on a more comprehensive definition of what we are to understand as correlation in the case of a complex of variables.

---

* *On the Theory of Contingency.* "Drapers' Research Memoirs, Biometric Series I" (Dulau & Co., Soho Square, London.

These points will be considered in the present treatment of correlation.

(2) *On Difference Methods of finding the Coefficient of Correlation.*

Let x and y be two correlated variables, each measured from their means m_1 and m_2 respectively. Then if $v=x-y$, and σ_x, σ_y, σ_v denote the three standard deviations

$$\sigma_v^2=\sigma_x^2+\sigma_y^2-2r_{xy}\sigma_x\sigma_y$$

and

$$r_{xy}=(\sigma_x^2+\sigma_y^2-\sigma_v^2)/(2\sigma_x\sigma_y) \quad \text{.............................(i).}$$

This method of finding r_{xy} has long been in use as an alternative method to the product-moment method†.

It involves finding the mean-square difference of the values of the pairs of correlated characters. It is possible, however, to find r_{xy} from about one half these differences, if we assume the distribution to be normal.

More generally I proceed as follows. Suppose the function $mx-ny$ formed, where m and n are at present indeterminate positive constants, and let the positive values only of this expression be taken and divided by the total frequency N. Then it will be possible to determine r from this result.

If z be the ordinate of the surface, then :

$$z=\frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_x^2}-\frac{2rxy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\right)} \quad \text{......................(ii)}$$

and we have the above result expressed analytically:

$$\frac{S(mx-ny)}{N}=\frac{1}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1-r^2}}\iint(mx-ny)\,e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_x^2}-\frac{2rxy}{\sigma_x\sigma_y}+\frac{y^2}{\sigma_y^2}\right)}dy\,dx \quad \text{...(iii).}$$

The limits to y in order that $mx-ny$ may be positive are $y=-\infty$ to mx/n, and the limits of x will then be $x=\infty$ to $-\infty$.

Put

$$y'=y/\sigma_2 \qquad x'=x/\sigma_1,$$

then

$$\frac{S(mx-ny)}{N}=\frac{1}{2\pi}\frac{1}{\sqrt{1-r^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{\frac{m\sigma_1}{n\sigma_2}x'}(m\sigma_1x'-n\sigma_2y')\times e^{-\frac{1}{2}\frac{1}{1-r^2}(x'^2-2rx'y'+y'^2)}dy'\,dx'.$$

---

* *Phil. Trans.* A, Vol. 200, pp. 1—66.

† For example, *Phil. Trans.* A, Vol. 198, p. 242, and often elsewhere.

Write
$$y'' = \frac{m\sigma_1}{n\sigma_2}\, x' - y',$$
and we have:
$$\frac{S\,(mx-ny)}{N} = \frac{1}{2\pi}\,\frac{n\sigma_2}{\sqrt{1-r^2}}\int_{-\infty}^{+\infty}\int_0^{\infty} y'' e^{-\frac{1}{2}\frac{1}{1-r^2}\left\{x'^2 - 2rx'\left(\frac{m\sigma_1}{n\sigma_2}x' - y''\right) + \left(\frac{m\sigma_1}{n\sigma_2}x' - y''\right)^2\right\}}\, dy''dx'.$$

The order of integration can now be changed and we have:
$$\frac{S\,(mx-ny)}{N} = \frac{1}{2\pi}\,\frac{n\sigma_2}{\sqrt{1-r^2}}\int_0^{\infty} y''\, e^{-\frac{1}{2}\frac{y''^2}{\alpha^2}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\frac{1}{\beta^2}(x'-\gamma y'')^2}\, dx'\, dy'',$$
where if $\epsilon = m\sigma_1/(n\sigma_2)$,
$$\frac{1}{\alpha^2} = \frac{1 - 2r\epsilon\,(1-\epsilon) - r^2\epsilon^2}{(1 - 2r\epsilon + \epsilon^2)\,(1-r^2)}, \qquad \frac{1}{\beta^2} = \frac{1 - 2r\epsilon + \epsilon^2}{1-r^2},$$
$$\gamma = \epsilon\,(1-r)/(1 - 2r\epsilon + \epsilon^2).$$

But the integral with regard to x' is $\sqrt{2\pi}\,\beta$, and
$$\int_0^{\infty} y''\, e^{-\frac{1}{2}\frac{y''^2}{\alpha^2}}\, dy'' = \alpha^2.$$

Hence:
$$\frac{S\,(mx-ny)}{N} = \frac{1}{2\pi}\,\frac{n\sigma_2}{\sqrt{1-r^2}}\sqrt{2\pi}\,\beta\alpha^2.$$

Or, for the positive summation
$$\frac{S\,(mx-ny)}{N} = \frac{n^2\sigma_2^2}{\sqrt{2\pi}}\,\frac{(1-r^2)\sqrt{n^2\sigma_2^2 - 2rmn\sigma_1\sigma_2 + m^2\sigma_1^2}}{n^2\sigma_2^2 - 2rm\sigma_1\,(n\sigma_2 - m\sigma_1) - r^2m^2\sigma_1^2} \quad \text{............(iv)}.$$

This general value does not appear to be likely to be of much service. If we take $m = n = 1$, we obtain the result of simply summing the positive differences of paired variates. It is:
$$\frac{S\,(x-y)}{N} = \frac{\sigma_2^2}{\sqrt{2\pi}}\,\frac{(1-r^2)\sqrt{\sigma_2^2 - 2r\sigma_1\sigma_2 + \sigma_1^2}}{\sigma_2^2 - 2r\sigma_1\,(\sigma_2 - \sigma_1) - r^2\sigma_1^2} \quad \text{..................(v)}.$$

(v) leads to an equation of the 5th order to find r and again does not appear to be likely to be of any service. The variates must be reduced to a common unit before they are handled if we are to make (iv) workable. Such a unit is the standard deviation.

If we write $m = \frac{1}{\sigma_1}$ $n = \frac{1}{\sigma_2}$, we have at once:
$$\frac{S\left(\frac{x}{\sigma_1} - \frac{y}{\sigma_2}\right)}{N} = \sqrt{\frac{1-r}{\pi}}.$$

Thus we find:
$$r = 1 - \pi\left\{\frac{S\left(\frac{x}{\sigma_1} - \frac{y}{\sigma_2}\right)}{N}\right\}^2 \quad \text{..............................(vi)}.$$

(vi) is an extremely neat formula and might be taken as the definition of a quantity measuring correlation. But the actual determination of correlation in this way, *i.e.* the reduction of each variate to a deviation from its mean measured in terms of its S. D. as unit, would probably be as troublesome as using the product-moment method.

One special case occurs, however, in which the above formula may possibly be of good service. Suppose the two variates have the same mean $=m$ and the same S. D. $=\sigma$, then:

$$r=1-\frac{\pi\{S(x-y)\}^2}{N^2\sigma^2}=1-\pi\frac{\{S(m+x-\overline{m+y})\}^2}{N^2\sigma^2} \quad \ldots\ldots\ldots\ldots\ldots\text{(vii)}.$$

Or, the coefficient of correlation is the result of subtracting from unity π times the square of the mean sum of the positive differences of paired variates divided by their common standard deviation.

For cases in which both variates are the same, brothers, cousins of the same sex, homotypes, etc., and especially for some cases of short series, the method may be of value.

Illustration I. Resemblance of Length of Little Finger in Male Cousins. I take a short series of 68 male pairs of cousins. The average value of the length measured on the little finger was 51·02 mm. and its standard deviation 2·721 mm. There were 33 positive differences of finger length giving $S(x-y)=87·6$ mm. Hence we had:

$$r=1-\pi\frac{(87·6/68)^2}{7·4036}=·296.$$

Found by the product-moment method the answer was ·287; the difference is well within the probable error of the latter value. The process of taking differences and summing was considerably shorter than finding a product moment.

Illustration II. Assortative Mating in the case of Paramecium. I take Dr Pearl's Table AA 3 from Vol. V. p. 295 of *Biometrika* for the lengths of conjugating Paramecia.

I choose this purposely because there was no difficulty above about the male cousins; there were only two equalities, the actual measurements of each individual being recorded. But in an ordinary correlation table owing to the method of grouping there will be a very considerable number of ties, and the problem arises how are they to be distributed. Clearly one half of them will be excesses and one half defects, if we suppose the odds against an actual tie in measuring to any degree of accuracy to be very great. Hence we may say that half the diagonal total is to be treated as in excess. But at what portion of the base unit are we to set the pair apart? If the frequency was uniformly distributed over the diagonal cells, we should take the average interval between a pair to be $\frac{1}{2}$ the base unit. But the material is almost always clustered inside the cell, and clearly $\frac{1}{2}$ is too much. The actual value to be taken would depend upon the value of the correlation and the size of the base unit. In fact we can only take a rough approximation. I suggest that $\frac{1}{3}$ will be found to work fairly well. Accordingly we take $\frac{1}{6}$ of the contents of the diagonal cells, multiplied by the base unit. The whole process may now be written as follows:

$\frac{1}{6}$	1	2	3	4	5	6	7	8	9	10	11	12
0	1	1	1	0	$\boxed{1}$	0	–	–	–	–	–	–
2	1	0	0	$\boxed{0}$	1	0	–	–	–	–	–	
4	3	4	$\boxed{1}$	0	0	1	–	–	–	–		
4	14	$\boxed{7}$	5	1	1	0	–	–	–			
30	$\boxed{25}$	9	5	1	1	0	–	–				
$\boxed{22}$	16	5	5	0	0	0	–					
10	16	7	1	2	0	1				16·3		
16	4	2	0	0	0	$\overline{2}$				85		
4	5	3	0	0	$\overline{4}$					76		
4	0	0	0	$\overline{4}$						54		
2	0	0	$\overline{18}$							16		
0	0	$\overline{38}$								20		
0	$\overline{85}$									12		
$\overline{98}$												

$$\sigma = 19{\cdot}112^*, \qquad S(x-y) = \overline{279{\cdot}3} \times 10$$

$$r = 1 - \pi\left(\frac{2793}{40 \times 19{\cdot}112}\right)^2 = {\cdot}581.$$

The value obtained by the product-moment method is ·588 ± ·022. The correlation Table is as follows:

Length of First Conjugant.

Length of Second Conjugant	160–9	170–9	180–9	190–9	200–9	210–9	220–9	230–9	240–9	250–9	260–9	270–9	280–9	Totals
160–9	—	1	1	1	—	1	—	—	—	—	—	—	—	4
170–9	1	2	1	—	—	—	1	—	—	—	—	—	—	5
180–9	1	1	4	3	4	1	—	—	1	—	—	—	—	15
190–9	1	—	3	4	14	7	5	1	1	—	—	—	—	36
200–9	—	—	4	14	30	25	9	5	1	1	—	—	—	89
210–9	1	—	1	7	25	22	16	5	5	—	—	—	—	82
220–9	—	1	—	5	9	16	10	16	7	1	2	—	1	68
230–9	—	—	—	1	5	5	16	16	4	2	—	—	—	49
240–9	—	—	1	1	1	5	7	4	4	5	3	—	—	31
250–9	—	—	—	—	1	—	1	2	5	4	—	—	—	13
260–9	—	—	—	—	—	—	2	—	3	—	2	—	—	7
270–9	—	—	—	—	—	—	—	—	—	—	—	—	—	0
280–9	—	—	—	—	—	—	1	—	—	—	—	—	—	1
Totals	4	5	15	36	89	82	68	49	31	13	7	0	1	400

---

* Pearl, *loc. cit.* p. 226, Table II.

we proceed thus: Read each column down to and including the diagonal cell, and place the total under the corresponding differences in the previous scheme. For example, take the sixth column; 1, 0, 1, 7, 25, 22, are the corresponding frequencies, and these numbers will be found, sloping from the column marked 5, *i.e.* difference 5 × 10, diagonally across the scheme. In this manner the columns of the table can be disposed in the scheme at once. The scheme columns are then added up and multiplied by the difference at the top, and, if multiplied again by the base unit, in this case 10, the total gives $S(x-y)$. The whole can be done with very great rapidity, and the correlation found in about 10 minutes if σ be known.

As other comparisons I give the homotypic results:

	Difference method	Product method
Monmouthshire Ashes (65,000)	·432	·405 ± ·011
Papaver Rhoeas (Quantocks) (19,790)	·523	·533 ± ·013
Ditto (Chilterns' Base) (25,160)	·395	·400 ± ·012

These results show that there exists quite a reasonable amount of agreement between the two methods, and the difference method is much the shorter when the table contains thousands of observations as in these cases. At the same time too much reliance must not be placed upon the difference method, not only because it assumes normality of distribution but because it involves a somewhat rough method of approximation in the case of the diagonal cell.

One further point may be noted. Suppose that rank in a series was a true character which could be dealt with by a difference formula like the above then r the correlation of the ranks would be given by

$$r = 1 - \pi \left\{ \frac{S(x-y)}{N\sigma} \right\}^2.$$

Now for such ranks $\sigma^2 = \frac{1}{12}(N^2-1)$, therefore

$$r = 1 - \frac{12\pi \{S(x-y)\}^2}{N^2(N^2-1)} \quad \text{...............................(viii).}$$

Dr Spearman has introduced a quantity R which he terms a "correlational coefficient*" and which he defines without any special justification by:

$$R = 1 - \frac{S(x-y)}{\frac{1}{6}(N^2-1)} \quad \text{.................................(ix).}$$

We should thus have:

$$1 - r = \frac{\pi(N^2-1)}{3N^2}(1-R)^2 \quad \text{...............................(x),}$$

which would give approximately: $r = 2R - R^2$.

---

[*] *Journal of Psychology*, Vol. II. p. 96.

This is, of course, not true, for the distribution of ranks is not normal; the exact formula will be given later; but it suffices to indicate that the actual distribution assumed for x and y will much influence the relation between r and R. Dr Spearman from trial gives the empirical formula

$$r = \sin\left(\frac{\pi}{2} R\right) \quad \text{.................................(xi)},$$

which is also incorrect. But the above relation shows that we are not *à priori* compelled to suppose that r merely changes its sign, not its numerical value when R changes sign.

(3) *On the Correlation of Grades.* A method of representing frequency has been introduced by Francis Galton in which the extent of variation of a character is expressed by the position of the individual bearing this character in the population. This method was originally spoken of as that of percentiles but more recently as that of grades. A fundamental feature of the method is that the grade is looked upon as an index to the variate, it is not considered as in itself significant, or treated as an independent character of the individual. In order, however, to pass from the grade to the variate it is absolutely necessary to make some hypothesis as to the nature of the distribution. The hypothesis hitherto made is that the frequency follows, at least fairly closely, the normal or Gaussian law. On this assumption, tables of the probability integral enable us to pass at once from the grade to the magnitude of the variate, and *vice versâ.* Quite recently, however, Dr Spearman has proposed that rank in a population for any variate should be considered as in itself the quantitative measure of the character, and he proceeds to correlate ranks as if they were quantitative measures of character, without any reference to the true value of the variate. This seems to me a retrograde step; hitherto we have dealt with grade or rank (I will distinguish between them presently) as an index to the variate, and to make rank into a unit itself cannot fail, I believe, to lead to grave misconception. Between mediocrities the unit of rank treated as a measure of a variate is practically zero, between extreme individuals, it is very large indeed. To state that two individuals differ by m ranks carries no meaning at all unless we add, (i) the size of the population dealt with, (ii) the position in the population of one or both individuals, and (iii) the nature of the frequency distribution which governs the population. I cannot therefore look upon the correlation of ranks as conveying any real idea of the correlation of variates, unless we have a means of passing from the correlation of ranks to the value of the correlation of the variates, i.e. the correlation of ranks can only be treated as a step subsidiary to determining the true variate correlation.

The correlation between variates can be made to change widely by preserving the same system of ranks, but by altering the nature of the frequency distribution. Thus consider the system:

Variates

x	-2	-1	$+1$	$+2$
y	-2	-1	$+1$	$+2$

Ranks

1	2	3	4
1	2	3	4

The correlation of variates is perfect and the correlation of ranks is also perfect. But we may also have:

Variates

x	-2	$-1{\cdot}9$	$+1{\cdot}9$	$+2$
y	-2	$-{\cdot}01$	$+{\cdot}01$	$+2$

Ranks

1	2	3	4
1	2	3	4

The correlation of variates is now ·72, but the correlation of ranks remains perfect and would indicate nothing of this great difference. I think that it is safe to assert that until some assumption is made, at least as to the approximate nature of the distribution, we cannot hope to avoid misconceptions if we use the method of ranks without reference to the rank as index of the variate.

In such a case there can hardly be a doubt that the best method is first to consider to what results normal distribution will lead us, and secondly if the formulae found turn out to be of a simple character to adopt these as the basis by definition of the variate correlation constant as found from a method of ranks. This will be the course adopted in the present memoir.

(4) Let there be a population of N members and let these be under investigation for two correlated characters, means m_1, m_2, standard deviations σ_1, σ_2, correlation r. I shall suppose normality of distribution. Let m_1+x, m_2+y be the deviations of the two characters in any individual. Then I term:

$$\left.\begin{aligned} g_1 &= \tfrac{1}{2}N + \frac{N}{\sqrt{2\pi}\,\sigma_1}\int_0^x e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}\,dx \\ g_2 &= \tfrac{1}{2}N + \frac{N}{\sqrt{2\pi}\,\sigma_2}\int_0^y e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}}\,dy \end{aligned}\right\} \qquad \text{(xii)},$$

the x- and y-*grades* of the variates for the individual. It will be obvious that g_1 and g_2 are mathematical functions of the variates and that accordingly the correlation between them determines that between x and y, or *vice versâ*.

Obviously g_1 and g_2 can be found from tables of the probability integral as soon as x and y, the deviates, are known.

I term *rank* the actual position in order of an individual with regard to any variate in a given series obtained by measurement or observation. If ν_1 be the 'rank' of an individual for a given character this signifies that in the observed population there are $\nu_1 - \frac{1}{2}$ individuals with character greater than x. If therefore we were to identify this with the grade we should have

$$g_1 = \nu_1 - \tfrac{1}{2} \quad \text{.......................................(xiii)},$$

or g_1 would always differ from a whole number by ·5. This, of course, it does not, and the whole problem of working with ranks really centres on the degree of approximation which is made when we proceed from ranks to grades by the relation (xiii). A grade determined from a rank and not from a variate we may term a spurious grade; actually the real grade often differs by several units from the spurious grade, and the practical problem is: To what extent does this vitiate the use of ranks as a subsidiary stage to the determination of variate-correlation?

I shall first proceed to find the mean and standard deviation of a true grade; (xii) shows us at once that $\bar{g}_1 = \bar{g}_2 = \frac{1}{2}N$ is the mean value of the grade.

The frequency of a given variate lying between x and $x + \delta x$

$$= \frac{N}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} dx = dg_1.$$

But the frequency of the variate must also be the frequency of its grade, or:

$$N\sigma_{g_1}^{\ 2} = \int_{-\infty}^{+\infty} (g_1 - \bar{g}_1)^2 \, dg_1 = \tfrac{1}{3}\Big[(g_1 - \bar{g}_1)^3\Big]_0^N$$

$$= \frac{2}{3}\frac{N^3}{8} = \frac{N^2}{12} \times N.$$

Hence we have: $$\sigma_{g_1}^{\ 2} = \sigma_{g_2}^{\ 2} = \tfrac{1}{12}N^2 \quad \text{.................................(xiv)}.$$

Now whereas our grades are a continuous series, the spurious grades or ranks are discontinuous and at intervals $h = 1$. (xiii) shows us at once that

$$\bar{\nu}_1 = \bar{\nu}_2 = \bar{g}_1 + \tfrac{1}{2} = \tfrac{1}{2}(N + 1).$$

Further $$\sigma_{g_1}^{\ 2} = \sigma_{\nu_1}^{\ 2} + \tfrac{1}{12}h^2,$$

the latter corresponding to the Sheppard's correction by which we pass from raw to adjusted moments.

Thus we have: $$\left.\begin{aligned} \sigma_{\nu_1}^{\ 2} = \sigma_{\nu_2}^{\ 2} &= \tfrac{1}{12}(N^2 - 1) \\ \bar{\nu}_1 = \bar{\nu}_2 &= \tfrac{1}{2}(N + 1) \end{aligned}\right\} \quad \text{............................(xv)},$$

(xv) must be used whenever we are dealing with ranks or spurious grades.

Writing $i_1 = g_1 - \bar{g}_1$ and $i_2 = g_2 - \bar{g}_2$, I now turn to the determination of the product moment of the grades. Let us put:

$$z = \frac{1}{2\pi\sigma_1\sigma_2}\frac{1}{\sqrt{1 - r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)} \quad \text{..............(xvi)},$$

then: $$p_{g_1 g_2} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1 i_2 z \, dx \, dy$$

gives the product moment of the grades.

Differentiate $p_{g_1g_2}$ with regard to r which is not contained in either i_1 or i_2; we have:

$$\frac{dp_{g_1g_2}}{dr} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1 i_2 \frac{dz}{dr}\, dx\, dy.$$

But I have elsewhere* shown that:

$$\frac{dz}{dr} = \sigma_1\sigma_2 \frac{d^2z}{dx\,dy} \quad\text{.................................(xvii).}$$

Accordingly:
$$\frac{dp_{g_1g_2}}{dr} = \sigma_1\sigma_2 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1 i_2 \frac{d^2z}{dx\,dy}\, dx\, dy.$$

Integrating twice by parts and noting that the part between limits vanishes in both cases, we have:

$$\frac{dp_{g_1g_2}}{dr} = \sigma_1\sigma_2 \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} z \frac{di_1}{dx}\frac{di_2}{dy}\, dx\, dy.$$

Substituting for di_1/dx and di_2/dy and writing $x = x'\sigma_1$ $y = y'\sigma_2$, we find:

$$\frac{dp_{g_1g_2}}{dr} = \frac{N^3}{4\pi^2\sqrt{1-r^2}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2(1-r^2)}\left\{(2-r^2)x'^2 - 2rx'y' + (2-r^2)y'^2\right\}}\, dx'\, dy'$$

$$= \frac{N^3}{2\pi\sqrt{1-r^2}} \frac{1}{\sqrt{\left(\frac{2-r^2}{1-r^2}\right)^2 - \frac{r^2}{(1-r^2)^2}}} = \frac{N^3}{2\pi\sqrt{4-r^2}}.$$

Now if ρ_{12} be the correlation of grades, we have:

$$\rho_{12} = \frac{p_{g_1g_2}}{N\sigma_{g_1}\sigma_{g_2}}, \text{ and } \frac{d\rho_{12}}{dr} = \frac{1}{N\sigma_{g_1}\sigma_{g_2}} \frac{dp_{g_1g_2}}{dr}.$$

Thus remembering (xiv)

$$\frac{d\rho_{12}}{dr} = \frac{6}{\pi}\frac{1}{\sqrt{4-r^2}},$$

or,
$$\rho_{12} = \frac{6}{\pi}\sin^{-1}\tfrac{1}{2}r + \text{constant}.$$

Now ρ_{12} and r must vanish together, hence the constant is zero. Accordingly we have:

$$r = 2\sin\left(\frac{\pi}{6}\rho_{12}\right) \quad\text{.................................(xviii).}$$

This remarkably simple formula enables us to determine the value of the true variate correlation from a correlation of grades on the assumption of the normal law; or if grades may be replaced by ranks, a knowledge of the correlation of ranks will give us the correlation of the actual variates behind the order exhibited in the ranking. The important idea embodied in the above formula is the basis of the present memoir, and is as far as I am aware wholly new.

---

* *Phil. Trans.* A. Vol. 195, p. 25.

It remains for us to consider methods of finding the rank or grade correlation and the probable error of such methods.

(5) A convenient method of finding the grade correlation is that of formula (i), p. 4, we have at once:

$$\rho_{12} = \frac{\sigma_1^2 + \sigma_2^2 - S(g_1 - g_2)^2/N}{2\sigma_1\sigma_2}.$$

Or,

$$\rho_{12} = 1 - \frac{6S(g_1 - g_2)^2}{N^3} \quad \text{..............................(xix)},$$

if we use true grades,

but:

$$= 1 - \frac{6S(\nu_1 - \nu_2)^2}{N(N^2 - 1)} \quad \text{..................................(xx)},$$

if we use ranks ν_1 and ν_2.

If we use ranks the discovery of $S(\nu_1 - \nu_2)^2$ or the sum of the squares of the differences of ranks forms a very easy process of determining ρ_{12}, due regard being paid to certain points to be dealt with in the illustrations below. Then (xviii) will give the variate correlation.

The probable error of ρ_{12} and of r found in this way will be given in another section.

Since the determination of ρ_{12} by (xx) is algebraically identical with finding ρ_{12} by the product moment, and such product moment gives the least probable error in the determination of a correlation coefficient, there must be some fallacy in a statement which has been propounded among the psychologists that a difference method of determining the correlation will give ρ_{12} with about $\frac{2}{3}$ of the probable error of the product moment method. This fallacy will be considered later.

Meanwhile it is of interest to show that the probable error* of

$$\rho_{12} = \{S(\nu_1\nu_2)/n - \bar{\nu}_{12}\bar{\nu}\}/(\sigma_{\nu_1}\sigma_{\nu_2})$$

is of the form:

$$\text{P.E.} = \frac{\cdot 67449}{\sqrt{n-1}}\left(1 - c_2\,\rho_{12}^2 + c_4\,\rho_{12}^4 + c_6\,\rho_{12}^6 + \ldots\right)$$

where $c_2, c_4, c_6, \ldots$ are undetermined constants. Or, the probable error of ρ_{12} for $\rho_{12} = 0$, or for uncorrelated ranks is

$$\cdot 67449/\sqrt{n-1},$$

i.e. is absolutely identical with probable error of a coefficient of correlation of any two uncorrelated variables, and is not as asserted much smaller.

Since for ranks σ_{ν_1} and σ_{ν_2} are constant, we have† to find the value of

$$u^2 = \Sigma\left\{\frac{S(\nu_1\nu_2)}{n} - \{\tfrac{1}{2}(n+1)\}^2\right\}^2,$$

ν_1 and ν_2 being independent, in order to reach the squared standard deviation of ρ_{12} for $\rho_{12} = 0$.

---

* n is here put for N as more convenient for the algebraic work which follows

† I owe the following proof to the kindness of my friend "Student."

Now: $$u^2 = \Sigma\left\{\left(\frac{S(\nu_1\nu_2)}{n}\right)^2 - 2\left(\frac{n+1}{2}\right)^2\frac{S(\nu_1\nu_2)}{n} + \left(\frac{n+1}{2}\right)^4\right\}.$$

There being no correlation, $n!$ arrangements of this product occur with equal frequency. Hence

$$\Sigma\left(\frac{n+1}{2}\right)^4 = \left(\frac{n+1}{2}\right)^4 n!.$$

Next any $\nu_1\nu_2$ occurs in $(n-1)!$ of the arrangements, for if ν_1 be paired with ν_2, the remaining $n-1$ pairs may be arranged in $(n-1)!$ ways. Thus

$$\Sigma\left\{2\left(\frac{n+1}{2}\right)^2 S(\nu_1\nu_2)/n\right\} = 2(n-1)!\left(\frac{n+1}{2}\right)^2\frac{\Sigma(\nu_1\nu_2)}{n}$$

$$= 2(n-1)!\left(\frac{n+1}{2}\right)^2\frac{\Sigma(\nu_1)\,\Sigma(\nu_2)}{n} = 2(n-1)!\left(\frac{n+1}{2}\right)^2\left\{\frac{n(n+1)}{2}\right\}^2\Big/ n$$

$$= 2(n)!\left(\frac{n+1}{2}\right)^4.$$

Further: $$\Sigma\left(\frac{S(\nu_1\nu_2)}{n}\right)^2 = \Sigma\frac{1}{n^2}\{S(\nu_1^2\nu_2^2) + 2S(\nu_1\nu_2\nu_1'\nu_2')\},$$

where ν_1', ν_2' are different from ν_1, ν_2.

Now $\nu_1^2\nu_2^2$ occurs in $(n-1)!$ arrangements; hence

$$\Sigma\left(\frac{S(\nu_1^2\nu_2^2)}{n^2}\right) = \frac{(n-1)!}{n^2}\Sigma(\nu_1^2)\,\Sigma(\nu_2^2) = \frac{(n-1)!}{n^2}\left(\frac{n(n+1)(2n+1)}{6}\right)^2$$

$$= (n-1)!\left\{\frac{(n+1)(2n+1)}{6}\right\}^2.$$

Next $\nu_1\nu_2\nu_1'\nu_2'$ occurs in $(n-2)!$ arrangements.

Thus: $$\Sigma\left(\frac{2S(\nu_1\nu_2\nu_1'\nu_2')}{n^2}\right) = \frac{2(n-2)!}{n^2}\Sigma(\nu_1\nu_2\nu_1'\nu_2')$$

$$= \frac{(n-2)!}{n^2}\Sigma(\nu_1\nu_2)\{\Sigma(\nu_1'\nu_2') - \nu_1\Sigma(\nu_2') - \nu_2\Sigma(\nu_1') + \nu_1\nu_2\},$$

where ν_1' and ν_2' may now take all values.

Thus:

$$\Sigma\left\{\frac{2S(\nu_1\nu_2\nu_1'\nu_2')}{n^2}\right\} = \frac{(n-2)!}{n^2}\left\{\Sigma(\nu_1\nu_2)\,\Sigma(\nu_1'\nu_2') - \frac{n(n+1)}{2}S(\nu_1^2\nu_2 + \nu_2^2\nu_1) + S(\nu_1^2\nu_2^2)\right\}$$

$$= \frac{(n-2)!}{n^2}\left\{\left(\frac{n(n+1)}{2}\right)^4 - 2\left(\frac{n(n+1)}{2}\right)^2\frac{n(n+1)(2n+1)}{6} + \left(\frac{n(n+1)(2n+1)}{6}\right)^2\right\}$$

$$= \frac{(n-1)!}{144}(n+1)^2\{9n^3 + 3n^2 - 8n - 4\}.$$

Collecting the various parts we find:

$$u^2 = \frac{(n-1)!\,(n+1)^2\,(2n+1)^2}{36} + \frac{(n-1)!}{144}(n+1)^2(9n^3+3n^2-8n-4)$$

$$-2n!\left(\frac{n+1}{2}\right)^4 + n!\left(\frac{n+1}{2}\right)^4,$$

or, after reducing,
$$u^2 = \frac{n!\,(n+1)^2\,(n-1)}{144}.$$

Therefore the mean value of u^2 is $\frac{(n+1)^2(n-1)}{144}$.

Now the probable error of ρ_{12}, for uncorrelated ranks:

$$= \cdot 67449\,u/\sigma_{\nu_1}\sigma_{\nu_2}.$$

$$= \cdot 67449\,\frac{(n+1)\sqrt{n-1}}{12}\Big/\frac{n^2-1}{12}$$

$$= \cdot 67449/\sqrt{n-1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xxi)}.$$

It thus follows that if the value of ρ_{12} be not two or three times the expression (xxi), there is no significant correlation of ranks, and therefore no significant correlation of the corresponding variates.

(6) *On the Difference Method of finding the Correlation of Grades.*

Exactly as in the first section of this paper we may seek the correlation of grades by means of the sum $S(g_1-g_2)$ of all their positive differences. This is slightly shorter than finding $S(g_1-g_2)^2$, but only very slightly so, and it may be doubted whether the increased rapidity of working at all compensates for the decreased accuracy of the process. Still the result is interesting and throws considerable light on one or two allied points.

Let $G=S(g_1-g_2)$, where the sum S is for all x-grades which are greater than corresponding y-grades.

Let us put $x=\sigma_1 x'$, $y=\sigma_2 y'$, and write

$$j_v = \int_0^v e^{-\frac{1}{2}v^2}\,dv,$$

$$z = \frac{1}{2\pi}\,\frac{1}{\sqrt{1-r^2}}\,\frac{1}{\sigma_1\sigma_2}\,z'.$$

Then:
$$G = \frac{N}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\int_{-\infty}^{x'}(j_{x'}-j_{y'})\,z\,dy\,dx$$

$$= \frac{N^2}{(2\pi)^{\frac{3}{2}}\sqrt{1-r^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{x'}(j_{x'}-j_{y'})\,z'\,dy'\,dx'.$$

$$\frac{dG}{dr}=\frac{N^2}{(2\pi)^{\frac{3}{2}}}\int_{-\infty}^{+\infty}\int_{-\infty}^{x'}(j_{x'}-j_{y'})\,\frac{d}{dr}\left(\frac{z'}{\sqrt{1-r^2}}\right)dy'dx'$$

$$=\frac{N^2}{(2\pi)^{\frac{3}{2}}}\int_{-\infty}^{+\infty}\int_{-\infty}^{x'}(j_{x'}-j_{y'})\,\frac{d}{dx'\,dy'}\left(\frac{z'}{\sqrt{1-r^2}}\right)dy'dx'$$

$$=\frac{N^2}{(2\pi)^{\frac{3}{2}}\sqrt{1-r^2}}\int_{-\infty}^{+\infty}\left\{\left[(j_{x'}-j_{y'})\,\frac{dz'}{dx'}\right]_{-\infty}^{x'}+\int_{-\infty}^{x'}\frac{dj_{y'}}{dy'}\,\frac{dz'}{dx'}\,dy'\right\}dx'$$

$$=\frac{N^2}{(2\pi)^{\frac{3}{2}}\sqrt{1-r^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{x'}e^{-\frac{1}{2}y'^2}\left(-\frac{(x'-ry')}{1-r^2}\,z'\right)dy'dx'.$$

Put $y'=x'-y''$; then after rearranging:

$$\frac{dG}{dr}=-\frac{N^2}{(2\pi)^{\frac{3}{2}}(1-r^2)^{\frac{3}{2}}}\int_{-\infty}^{+\infty}\int_{0}^{\infty}\left\{(1-r)\left(x'-\frac{2+r}{3+r}\,y''\right)+\frac{2\,(1+r)}{3+r}\,y''\right\}$$

$$\times\, e^{-\frac{1}{2}\left\{\frac{3+r}{1+r}\left(x'-\frac{2+r}{3+r}y''\right)^2+\frac{2}{(1-r)(3+r)}y''^2\right\}}\,dy''dx'.$$

The order of integration can now be transposed and if X be written for

$$x'-\frac{2+r}{3+r}\,y'',$$

the limits of X will also be $-\infty$ to $+\infty$. Thus:

$$\frac{dG}{dr}=-\frac{N^2}{(2\pi)^{\frac{3}{2}}(1-r^2)^{\frac{3}{2}}}\int_{0}^{\infty}e^{-\frac{1}{2}\frac{2}{(1-r)(3+r)}y''^2}\int_{-\infty}^{+\infty}\left\{(1-r)\,X+\frac{2(1+r)}{3+r}\,y''\right\}e^{-\frac{1}{2}\frac{3+r}{1+r}X^2}\,dXdy''.$$

But if c have any value:

$$\int_{-\infty}^{+\infty}e^{-\frac{1}{2}c^2X^2}\,XdX=0,\quad\text{and}\quad\int_{-\infty}^{+\infty}e^{-\frac{1}{2}c^2X^2}\,dX=\sqrt{2\pi}\,\frac{1}{c}.$$

Hence:

$$\frac{dG}{dr}=-\frac{N^2}{(1-r^2)^{\frac{3}{2}}}\,\frac{1}{(2\pi)}\,\frac{2\,(1+r)}{3+r}\sqrt{\frac{1+r}{3+r}}\int_{0}^{\infty}y''e^{-\frac{1}{2}\frac{2}{(1-r)(3+r)}y''^2}\,dy''$$

$$=-\frac{N^2}{2\pi}\,\frac{1}{(1-r^2)^{\frac{3}{2}}}\,\frac{2\,(1+r)^{\frac{3}{2}}}{(3+r)^{\frac{3}{2}}}\,\frac{(1-r)\,(3+r)}{2}$$

$$=-\frac{N^2}{2\pi}\,\frac{1}{\sqrt{(1-r)\,(3+r)}}=-\frac{N^2}{2\pi}\,\frac{1}{\sqrt{4-(1+r)^2}}.$$

Hence integrating:

$$G=\text{constant}+\frac{N^2}{2\pi}\cos^{-1}\frac{1+r}{2}.$$

But when $r=1$, G must be zero; therefore the constant is zero, or inverting:

$$r=2\cos 2\pi\,\frac{G}{N^2}-1.$$

Or, finally*:
$$r = 2\cos 2\pi\left(\frac{S(g_1-g_2)}{N^2}\right) - 1 \quad\text{(xxii).}$$

This gives us the correlation of two variates from the corresponding grades by a difference method.

If r be zero, we must have $2\pi\frac{S(g_1-g_2)}{N^2}$ equal to $60° = \pi/3$, or $S(g_1-g_2) = \frac{1}{6}N^2$ when there is no correlation of variates. This is easily proved directed, for in this case:

$$G = S(g_x - g_y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{x'} (g_x - g_y)\frac{N}{2\pi}e^{-\frac{1}{2}(x'^2+y'^2)}\,dy'dx'$$
$$= \frac{1}{N}\int_0^N\int_0^{g_x}(g_x-g_y)\,dg_y\,dg_x$$
$$= \frac{1}{N}\int_0^N \frac{g_x^2}{2}\,dg_x = \frac{N^2}{6}.$$

For ranks the corresponding expression to be used is $\frac{1}{6}(N^2-1)$, or we have:

$$r = 2\cos 2\pi\left(\frac{S(\nu_1-\nu_2)}{N^2-1}\right) - 1 \quad\text{(xxiii).}$$

As before the truth of (xxii) depends on the approximation to normal correlation.

If we combine (xx), (xviii) and (xxiii) we have the relation between $S(\nu_1-\nu_2)^2$ and $S(\nu_1-\nu_2)$ which holds in the case of normal correlation.

Writing $R = 1 - S(\nu_1-\nu_2)/\frac{1}{6}(N^2-1)$, we have:

$$r = 2\sin\frac{\pi}{6}\rho_{12} = 2\cos\frac{\pi}{3}(1-R) - 1 \quad\text{(xxiv).}$$

Dr Spearman gives† the relation:

$$\rho_{12} = \sin\left(\frac{\pi}{2}R\right) \quad\text{(xxv)}$$

(he neither connects ρ_{12}, nor R, with r) as apparently an empirical relationship and speaks of it as "all that could be desired." It is clearly incompatible with normal

---

[*] The relationship of (xxii) to (vii) is easily seen if we expand the cosine as far as the square of the angle. We have

$$r = 1 - \frac{4\pi^2}{N^4}\{S(g_1-g_2)\}^2 = 1 - \frac{\pi}{3}\frac{\pi\{S(g_1-g_2)\}^2}{N^2\sigma^2} = 1 - 1{\cdot}0472\,\frac{\pi\{S(g_1-g_2)\}^2}{N^2\sigma^2}.$$

(vii) would have given us 1 instead of the factor 1·0472. Thus when there is high correlation, or $S(g_1-g_2)$ is small, we see that the difference method with grades leads us to nearly the same result, as the assumption that the grades themselves form a normal distribution. This suggests that Spearman would have got much better results for his "footrule" for measuring correlation had he taken $R = 1 - 3\frac{N^2}{N^2-1}\left(\frac{S(\nu_1-\nu_2)}{N\sigma}\right)^2$ instead of $1 - \frac{S(\nu_1-\nu_2)}{2\sigma^2}$; for this value, i.e. $1 - \left(\frac{S(g)}{M}\right)^2$ in his notation, would have been almost the true variate correlation r.

[†] *Journal of Psychology*, Vol. II. p. 102.

correlation, which at any rate is a fairly good guide for general relations of this sort in the theory of frequency. Table I. gives the values of r and R for each ·05 for ρ_{12}. Table II. gives the values of r and ρ_{12} for each ·05 of R, and in the last column the value of ρ_{12} which would arise if (xxv) were correct.

TABLE I. *Correlation of Variates from Mean Square Difference of Grades.*

ρ_{12}	r	R	ρ_{12}	r	R
·00	·000	·000	·50	·518	·323
·05	·052	·029	·55	·568	·361
·10	·105	·059	·60	·618	·400
·15	·157	·089	·65	·668	·442
·20	·209	·120	·70	·717	·486
·25	·261	·152	·75	·765	·533
·30	·313	·184	·80	·813	·584
·35	·364	·217	·85	·861	·644
·40	·416	·251	·90	·908	·709
·45	·467	·286	·95	·954	·796
·50	·518	·323	1·00	1·000	1·000

TABLE II. *Correlation of Variates from Difference of Grades.*

R	r	ρ_{12}	(xxv)	R	r	ρ_{12}	(xxv)
·00	·000	·000	·000	·50	·732	·716	·707
·05	·089	·085	·078	·55	·782	·767	·760
·10	·176	·168	·156	·60	·827	·814	·809
·15	·259	·248	·233	·65	·867	·856	·853
·20	·338	·324	·309	·70	·902	·894	·891
·25	·414	·398	·383	·75	·932	·926	·924
·30	·486	·469	·454	·80	·956	·952	·951
·35	·554	·536	·522	·85	·975	·973	·972
·40	·618	·600	·587	·90	·989	·988	·988
·45	·677	·660	·649	·95	·997	·997	·997
·50	·732	·716	·707	1·00	1·000	1·000	1·000

Now these Tables bring out several interesting facts. The first is the remarkable closeness between the correlation of the grades and the true correlation of the variates, if we suppose the system normal. The maximum difference as shown by Table I. is ·018 and actually the maximum of $r - \rho_{12}$ occurs when $\rho_{12} = {}$·5756 and is then ·0180. Thus, the difference will often be of the order of the probable error. The formula (xviii) is so simple, that we can always deduce the variate correlation at once from the grade correlation. I propose to define r as given by (xviii) as the *grade-variate correlation.* Whenever the system is normal, or approximately normal, this will agree with the true variate correlation closely. Next Table I. shows us that equal differences of ρ_{12} give almost equal differences of r, i.e. the differences only range from ·052 to ·046 of r for differences of ·050 of ρ_{12}. On the other hand the differences of r for equal differences ·050 of R vary from ·089 to ·003, or second differences become of importance. Clearly for high values of R, r will be found much more closely than for low values.

If E_r' be the error in r due to an error $E_{\rho_{12}}$ in ρ_{12}, and E_r'' be the error due to an error E_R in R, we have:

$$E_r' = \frac{\pi}{3}\cos\frac{\pi}{6}\rho_{12} \times E_{\rho_{12}},$$

$$E_r'' \doteq \frac{2\pi}{3}\sin\frac{\pi}{3}(1-R)\times E_R$$

if we use differentials. For the special case of $\rho_{12} = R = 0$, we have seen that the probable error of $\rho_{12} = \cdot67449/\sqrt{n-1}$; it will be seen later that the probable error of R is $\cdot4266/\sqrt{n-1}$ nearly, and if E_r be the probable error of $r=0$, as found in the ordinary product moment way, we have:

$$E_r : E_r' : E_r'' :: \frac{\cdot6745}{\sqrt{n-1}} : \frac{\pi}{3}\frac{\cdot6745}{\sqrt{n-1}} : \frac{2\pi}{3}\frac{\sqrt{3}}{2}\frac{\cdot4266}{\sqrt{n-1}}$$

$$:: \frac{\cdot6745}{\sqrt{n-1}} : \frac{\cdot7063}{\sqrt{n-1}} : \frac{\cdot7738}{\sqrt{n-1}} \quad \text{......................(xxvi).}$$

Thus we see that, contrary to what has been asserted, the accuracy of the new methods—when they are measured by the determination of the true correlation—are less than the old product moment method. In particular it requires about 30 per cent. more observations by the R method to obtain r with the same degree of certainty, when $r=0$.

At present we do not know the R factor term in E_R, when R differs from zero, and accordingly cannot test E_r, E_r' and E_r'' at other values of R or ρ_{12}, but I have little doubt of the general truth of the result that E_r is at all values as well as for $r=0$, sensibly less than E_r' and still less than E_r''.

(7) *Remarks on the Probable Error of R.*

The probable error of a quantity in which the limits of the summation vary as we make random variations in the constants is always a troublesome matter, and I have not yet succeeded in evaluating the probable error of $S(g_1-g_2)$ when $g_1>g_2$ for any value of r.

Spearman has investigated the probable error of the corresponding expression for ranks, $S(\nu_1-\nu_2)$, *when there is no correlation between the ranks.* He finds that for n observations the probable error of R may be taken as $\cdot43/\sqrt{n}$, and from this result he has drawn rather sweeping conclusions as that: "twenty cases treated in one of the ways described furnish as much certitude as 180 in another more usual way"; or that: "a probable error may at present be admitted without much hesitation up to

0·05; so that by adopting the method of calculation recommended, two to three dozen subjects would be sufficient for most purposes*." Now these statements seem to me not without grave danger, and accordingly it is well to see where the error has crept in.

Spearman gives the value $\cdot 4266/\sqrt{n}$, but it should be $\cdot 4266/\sqrt{n-1}$†, and accordingly since we have seen that the probable error of ρ_{12} for $\rho_{12}=0$, is $\cdot 6745/\sqrt{n-1}$, the probable error of R would only be about $\frac{2}{3}$ of the probable error of ρ_{12}, and upon this Spearman's statements are based.

Now the probable error of any quantity is conventionally $\cdot 67449 \times$ standard deviation$/\sqrt{n-1}$, and accordingly for the same number of observations the probable error is less when the standard deviation is less. But there would be no meaning in asserting that the mean of 20 metacarpal bones could be found with much more exactitude than the mean of 20 humeri, because the latter being a larger bone had a greater variability. We must either measure the same quantity by different processes, or else be at any rate certain that our quantities are alike in character and function before we compare their probable errors. The probable error of $\frac{1}{2}x$ is certainly less than that of x. Now ρ_{12} is a true correlation and ranges from $+1$ through 0 to -1 with a symmetrical distribution about 0, if we take the case of a random distribution of ranks. The quantity R presents nothing of this nature at all; random distribution of ranks does not give a symmetrical distribution for R, its range is not from $+1$ to -1, and there are certain values it can never take. In order to bring out these points I take the following table for R negative.

TABLE III. *Negative Correlation of Variates from Difference of Grades.*

R	r	ρ_{12}
−·05	−·092	−·088
−·10	−·187	−·178
−·15	−·283	−·271
−·20	−·382	−·367
−·25	−·482	−·465
−·30	−·584	−·566
−·35	−·687	−·670
−·40	−·791	−·777
−·45	−·895	−·886
−·50	−1·000	−1·000

N.B. It will be observed that when R is negative, the true variate correlation is almost double the magnitude of R, while if R be positive (Table II.) r is larger than R but not to this exaggerated extent. It will be clear that no estimate of the real correlation can be based on R, if it does not allow for this exaggeration.

---

[*] *American Journal of Psychology*, Vol. XV. pp. 100, 101. For the proof of the probable error cited see: *British Journal of Psychology*, Vol. II. pp. 105–8.

[†] Spearman's result at bottom of p. 108 may be written $\frac{\cdot 4266}{\sqrt{n-1}}\sqrt{\frac{n^2-3\cdot 5}{n^2-1}}$, or neglecting terms in $\frac{1}{n^2}$ *not* $\frac{1}{n}$ *as he does*, this gives $\cdot 4266/\sqrt{n-1}$ as we should anticipate.

Thus we see that while r and ρ_{12} run from $-0{\cdot}0$ to $-1{\cdot}0$, R only runs from $-0{\cdot}0$ to $-{\cdot}50$.

In order to obtain his probable error for R Spearman takes every random arrangement of ranks ν_1 and ν_2 for which ν_1 is greater than ν_2. He has neglected to observe that when he does this his R will become negative, but that it will not range from $-1{\cdot}0$ to $+1{\cdot}0$. For example, I take the following system of ranks for $(2m+1)$ individuals:

$\nu_1 =$	1	2	3	4	.	.	$2m+1$
$\nu_2 =$	$2m+1$	$2m$	$2m-1$	$2m-2$	.	.	1

This gives:
$$S(\nu_1-\nu_2)^2 = 2\{(2m)^2+(2m-2)^2+(2m-4)^2+\ldots+2^2\}$$
$$= 8m(m+1)(2m+1)/6,$$
$$\therefore\ \rho_{12} = 1-\frac{6S(\nu_1-\nu_2)^2}{N(N^2-1)} = 1-\frac{8m(m+1)(2m+1)}{(2m+1)((2m+1)^2-1)} = -1.$$

But
$$S(\nu_1-\nu_2) = 2+4+\ldots+(2m-4)+(2m-2)+2m = m(m+1).$$

Therefore:
$$R = 1-\frac{6S(\nu_1-\nu_2)}{N^2-1} = 1-\frac{6m(m+1)}{(2m+1)^2-1} = -{\cdot}5 \quad \ldots\ldots\ldots\ldots\ldots(\text{xxvii}).$$

Accordingly when the correlation is negative and perfect, the number of observations being odd, R will never take the value -1, but no greater value than $-{\cdot}5$; whereas if we reckon our second ranks in the negative direction R will equal $+1$.

Here the Spearman formula (xxv) leads to the absurd result $\rho_{12} = -1/\sqrt{2}$, instead of -1. On the other hand my formulae (xxiv) for $\rho_{12} = -1$ and $R = -{\cdot}5$ give absolutely the correct value $r = -1$ for the variate correlation.

Again take N even $= 2m$ and consider the system:

$\nu_1 =$	1	2	3	4	.	.	$2n$
$\nu_2 =$	$2m$	$2m-1$	$2m-2$	$2m-3$	.	.	1

We find:
$$S(\nu_1-\nu_2)^2 = 2\{(2m-1)^2+(2m-3)^2+(2m-5)^2+\ldots+1^2\}$$
$$= \frac{2m}{3}(4m^2-1),$$

and this gives
$$\rho_{12} = 1-\frac{6\,\frac{2m}{3}(4m^2-1)}{2m(4m^2-1)} = -1.$$

Again:
$$S(\nu_1-\nu_2) = 1+3+5+\ldots+(2m-5)+(2m-3)+(2m-1)$$
$$= m^2.$$

Hence:
$$R=1-\frac{6m^2}{4m^2-1}=-\frac{2m^2+1}{4m^2-1}=-\cdot 5\left\{1+\frac{3}{N^2-1}\right\} \quad\text{......(xxviii)}.$$

For:
$$N=4,\quad R=-\cdot 600;\quad N=10,\quad R=-\cdot 515,$$
$$N=20,\quad R=-\cdot 504;\quad N=100,\quad R=-\cdot 500.$$

Or, again, the limit $-\cdot 5$ is rapidly reached as the number of observations increases. In fact solely for the simple case of *two* observations is it possible for R to reach -1.

If it be objected to (xxiv) that it would now give for values of R, greater than $-\cdot 5$ values of the variate correlation greater than -1 ($=-1\cdot 09$ at a maximum for $N=4$), this is overlooking the point that (xxiv) is deduced from (xxii) by replacing true grades by spurious grades or ranks, and that if we retain (xxii) then

$$S(g_1-g_2)/N^2=\frac{m^2}{4m^2}=\tfrac{1}{4},$$

and $r=-1$ as it should do.

We have now reached I think the basis of Spearman's apparent paradox. While the variation of the true rank correlation ρ_{12} lies between $+1$ and -1 and has $\cdot 67449/\sqrt{N-1}$ for its probable error, the value of R only ranges between $+1$ and $-\cdot 5$, and may well have a less value for its probable error.

Now Spearman tells us that large negative values of his R should be avoided*. There is no necessity whatever for avoiding them if we are seeking the variate correlation by the formula given in this memoir. But if we are seeking the probable error of a zero quantity, which may vary on either side of zero (and in this case the variation is not symmetrical about zero), we cannot neglect the distribution of random variations below zero. If Spearman wishes his R to be considered always positive, then he ought to have found the probable error on the assumption that $S(\nu_1-\nu_2)$ should never be greater than $\frac{1}{6}(N^2-1)$. He has taken a quantity which ranges from $+1$ to $-\cdot 5$ and compared its random variations with one which ranges from $+1$ to -1 for the same frequency. If he had restricted his attention to variations of R between 0 and $+1$ and of ρ_{12} between 0 and $+1$ he would not have reached the same conclusion.

But there is a further very serious indictment to be made against Spearman's R. For values of N fairly small, which are those for which he proposes to use it, R retains a constant value for wide variations in ρ_{12}. We can show this on an exaggerated scale by writing down the possible values for Spearman's R and the true rank correlation for 4 individuals taken with random ranks. See Table on p. 23.

A little consideration will show to what much better results ρ_{12} leads us than R. R in fact remains constant and $=-\cdot 2$ while ρ_{12} passes through the values $0, -\cdot 2, -\cdot 4$ and $-\cdot 6$; or r can take values from 0 to $-\cdot 62$, while its value as found from R

---

* *Loc. cit.*, footnote, p. 96.

Correlation of Ranks and Variate Correlation for Four Individuals.

	ν_2																							
ν_1 1	1	1	1	2	2	1	1	2	3	1	3	2	3	2	4	2	3	4	4	3	3	4	4	4
2	2	2	3	1	1	3	4	3	1	4	2	4	1	3	1	4	2	1	2	4	4	2	3	3
3	3	4	2	3	4	4	2	1	2	3	1	1	4	4	2	3	4	3	1	1	2	3	1	2
4	4	3	4	4	3	2	3	4	4	2	4	3	2	1	3	1	1	2	3	2	1	1	2	1
R	+1·0	+·6	+·6	+·6	+·2	+·2	+·2	+·2	+·2	+·2	+·2	−·2	−·2	−·2	−·2	−·2	−·2	−·2	−·2	−·2	−·6	−·6	−·6	−·6
ρ_{12}	+1·0	+·8	+·8	+·8	+·6	+·4	+·4	+·4	+·4	+·2	+·2	0	0	−·2	−·2	−·4	−·4	−·4	−·4	−·6	−·8	−·8	−·8	−1·0
r from R	+1·0	+·83	+·83	+·83	+·34	+·34	+·34	+·34	+·34	+·34	+·34	−·38	−·38	−·38	−·38	−·38	−·38	−·38	−·38	−·38	−1·09	−1·09	−1·09	−1·09
r from ρ_{12}	+1·0	+·81	+·81	+·81	+·62	+·42	+·42	+·42	+·42	+·21	+·21	0	0	−·21	−·21	−·42	−·42	−·42	−·42	−·62	−·81	−·81	−·81	−1·0

I have again to thank "Student" for the values of R and ρ_{12}.

remains $-{\cdot}38$. This simple illustration of how the real rank correlation varies widely while Spearman's coefficient R remains constant shows how unsuitable the latter is, when we have to deal with small series.

Another point worth noting is that, if we take the positive values of the correlation only, the mean value of R is ·3818, while the mean value of the corresponding ρ_{12}'s is ·5454; the former has a standard deviation of ·2622 and the latter of ·2573, showing that we are not justified in asserting that R has a smaller probable error than ρ_{12} when we take comparable quantities.

Spearman appears to have an idea that R is really a coefficient comparable with ρ_{12}, and he attempts to get over some difficulties which have arisen, by telling us to reverse one series of ranks when R comes out negative. But reversing the ranks does not aid us to the right result. Thus if the ranks in the 12th and 13th column of ν_2 above be reversed, we find that R still remains negative and of the same magnitude $-{\cdot}2$. In fact it is easy to write down a system of ranks which give a negative R, and which on reversal give a negative R six or seven times as big. The fact is simply that R is not a symmetrical function of ρ_{12} and reversal of ranks does not necessarily reverse ρ_{12} in sign.

We see accordingly (i) that the total range of R is only about $\frac{2}{3}$ that of ρ_{12}, and that if we make the range the same by any attempt to reverse ranks, the Spearman method of calculating the probable error for $R=0$ is erroneous. (ii) That the distribution of R for random rankings has a median which differs from zero, is very skew, and is in no ways comparable with that for ρ_{12}.

A point to be borne in mind most carefully is that for a given value of R, ρ_{12} the true rank correlation may take a great variety of values. It is only when (i) the number of observations is fairly considerable, and (ii) we assume some distribution of associated grades such as that of normal correlation, that we are able to assert that the value of R will fix ρ_{12}, but such a relationship as that connecting ρ_{12}, R and the variate correlation r can only be fixed, as in this memoir, by the appeal to despised mathematical analysis.

Thus the advantages claimed by Spearman for R, namely: (a) that it frees the discussion from the complexities of mathematical analysis, and (b) that it gives a less probable error than more usual ways of approaching the subject, are seen to be illusory.

The difficulty that ρ_{12} may take a whole series of values for a single value of R is only surmounted if we define the character of our frequency distribution, and there is no doubt that we shall obtain a first approximation by defining it as normal. Secondly, we cannot reverse ranks with the effect Spearman proposes, and if we could his probable error of R for $R=0$ would be erroneous. Lastly, if we do not reverse ranks, then the probable error of one and the *same* quantity, the variate correlation, is considerably greater—for the only case yet worked out—i.e. $R=0$, when found by Spearman's method, than when found by the well-known method of squares of differences, and still less than if found by the product of the variates directly. The squares of the differences of ranks can be taken so directly and quickly from a table of squares, that it does not seem to me that the slight rapidity gained in using positive differences of ranks is of any weight against its increased inaccuracy for small series, where indeed it is likely to be chiefly used.

Further no two rank correlations are in the least reliable or comparable unless we assume that the frequency distributions are of the same general character (see p. 9), and this general character will, till further advance be made in the theory of skew-correlation, be undoubtedly that provided by the hypothesis of normal distribution. On this assumption Spearman's suggestion of correlation of ranks becomes valid, but not as he supposes as a *Ding an sich*, but only as a means of passing at any rate to an approximation to the variate correlation, and this in the case of quantities where it is easier to rank individuals than to measure their attributes accurately.

For the grounds stated in this section, I propose to use as a rule ρ_{12} and not R to find r. For this reason I have spent my energies in finding the probable error of ρ_{12} instead of seeking that of R.

(8) *On the Probable Error of the Correlation of Grades.*

The following investigation is admittedly lengthy, but I have not seen my way to shorten it, and the main point is to reach by some road the expression for the probable error. The most general expression for the probable error of a correlation whatever be the distribution is to be found from $\cdot 67449\Sigma_r$ where*:

$$\Sigma_r^{\ 2} = \frac{r^2}{N}\left\{\frac{p_{22}}{p_{11}^{\ 2}} + \frac{1}{2}\frac{p_{22}}{p_{20}\,p_{02}} + \frac{1}{4}\frac{p_{40}}{p_{20}^{\ 2}} + \frac{1}{4}\frac{p_{04}}{p_{02}^{\ 2}} - \frac{p_{31}}{p_{11}\,p_{20}} - \frac{p_{13}}{p_{11}\,p_{20}}\right\}$$

and

$$p_{qs} = S\{n_{xy}\,(x-\bar{x})^q\,(y-\bar{y})^s\}/N \text{ ....(xxix).}$$

Now in our case x and y are to be the grades g_1 and g_2 and r is to be

$$\rho_{12} = p_{11}/\sqrt{p_{20}\,p_{02}},$$

which we will write for this investigation ρ.

We have at once:

$$N \times p_{mo} = \int_{-\infty}^{+\infty} (g_1 - \bar{g}_1)^m \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}x^2/\sigma_1^2}\,dx,$$

and:

$$g_1 = \bar{g}_1 + \frac{N}{\sqrt{2\pi}\sigma_1}\int_o^x e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}\,dx,$$

$$dg_1 = \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}\,dx.$$

---

* *Mathematical Contributions to the Theory of Evolution*, XIV. "Draper's Research Memoirs," Biometric Series II, p. 20. I have omitted certain terms which cancel.

Thus:
$$p_{mo}=\frac{1}{N}\int_{o}^{N}(g_1-\bar{g}_1)^m\,dg_1=o,\ \text{if } m \text{ be odd,}$$
$$=\frac{1}{m+1}\left(\tfrac{1}{2}N\right)^m,\ \text{if } m \text{ be even.}$$

Thus
$$p_{40}=p_{04}=\tfrac{1}{80}N^4,\ \text{and}\ p_{20}=p_{02}=\tfrac{1}{12}N^2,\ \text{as before} \qquad \text{...........(xxx).}$$
We can now write (xxix) in the form:
$$\Sigma_\rho^{\ 2}=\frac{1}{N}\left\{\frac{p_{22}}{p_{20}^{\ 2}}\left(1+\tfrac{1}{2}\rho^2\right)-2\rho\,\frac{p_{31}}{p_{20}^{\ 2}}+\tfrac{9}{10}\rho^2\right\} \qquad \text{.................(xxxi),}$$
assuming as we shall show in the sequel (p. 30) that $p_{31}=p_{13}$. Accordingly we have now to find p_{22} and p_{31}.

First to find:
$$p_{22}=\frac{1}{N}\left\{S\,(g_1-\bar{g}_1)^2\,(g_2-\bar{g}_2)^2\right\},$$
or if we use the notation of p. 11,
$$p_{22}=\frac{1}{N}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1^{\ 2}\,i_2^{\ 2}\,z\,dx\,dy \qquad \text{......................(xxxii).}$$

Now I have not succeeded in integrating this expression, although I have spent much time over it, but I have expanded it in powers of the variate correlation r.

If
$$U=\frac{1}{\sqrt{1-r^2}}\,e^{-\frac{1}{2}\frac{1}{1-r^2}(x'^2-2rx'y'+y'^2)},$$
and $j_{x'}$ and $j_{y'}$ are the same as on p. 15, we can write:
$$p_{22}=\frac{N^4}{8\pi^3}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} j_{x'}^{\ 2}\,j_{y'}^{\ 2},\ U\,dx'dy'.$$

But*
$$U=S\,\frac{r^n}{n!}\,v_n w_n e^{-\frac{1}{2}(x'^2+y'^2)},$$
or,
$$p_{22}=\frac{N^4}{8\pi^3}\,S\left(\frac{r^n}{n!}\,q_n^{\ 2}\right),$$
where
$$q_n=\int_{-\infty}^{+\infty} j_{x'}^{\ 2}\,e^{-\frac{1}{2}x'^2}\,v_n\,dx'=\int_{-\infty}^{+\infty} j_{y'}^{\ 2}\,e^{-\frac{1}{2}y'^2}\,w_n\,dy' \qquad \text{...........(xxxiii).}$$

If n be odd, v_n and w_n have odd powers and $q_n=0$, hence p_{22} contains only even powers of r.

First:
$$q_0=\int_{-\infty}^{+\infty} j_x^{\ 2}\,e^{-\frac{1}{2}x^2}\,v_0\,dx,$$
where we may drop the dashes from the letters now, and $v_0=1$. Therefore:
$$q_0=\int_{-\infty}^{+\infty} j_x^{\ 2}\,\frac{dj_x}{dx}\,dx=\tfrac{1}{3}\left[\,j_x^{\ 3}\right]_{-\sqrt{\frac{1}{2}\pi}}^{+\sqrt{\frac{1}{2}\pi}}=\sqrt{\frac{2\pi}{3}}\times\frac{\pi}{\sqrt{12}} \qquad \text{.........(xxxiv).}$$

---

* Pearson: *Phil. Trans.* A, Vol. 195, p. 3.

Now*: $$v_n e^{-\frac{1}{2}x^2} = -\frac{d}{dx}(v_{n-1}e^{-\frac{1}{2}x}), \text{ and } v_{n-1} = \frac{1}{n}\frac{dv_n}{dx}.$$

Hence: $$q_n = -\int_{-\infty}^{+\infty} j_x^2\, d(v_{n-1}e^{-\frac{1}{2}x^2}) = 2\int_{-\infty}^{+\infty} j_x e^{-x^2} v_{n-1}\, dx$$

$$= \frac{2}{n}\int_{-\infty}^{\infty} j_x e^{-x^2}\, dv_n = -\frac{2}{n}\int_{-\infty}^{+\infty} v_n (e^{-\frac{3}{2}x^2} - 2xe^{-x^2} j_x)\, dx$$

$$= -\frac{2}{n}\int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2}\, dx + \frac{4}{n}\int_{-\infty}^{+\infty} j_x e^{-x^2} x v_n\, dx.$$

But†: $$xv_n = v_{n+1} + \frac{dv_n}{dx}, \text{ thus:}$$

$$q_n = -\frac{2}{n}\int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2}\, dx + \frac{4}{n}\int_{-\infty}^{+\infty} j_x e^{-x^2} v_{n+1}\, dx + \frac{4}{n}\int_{-\infty}^{+\infty} j_x e^{-x^2}\frac{dv_n}{dx}\, dx$$

$$= -\frac{2}{n}\int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2}\, dx + \frac{2}{n} q_{n+2} + 2q_n.$$

Or: $$q_{n+2} + \tfrac{1}{2} n q_n = \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2}\, dx = \beta_n \text{ say, } \quad \ldots\ldots\ldots\ldots\ldots\text{(xxxv)}.$$

It now remains to find β_n.

$$\beta_n = \int_{-\infty}^{+\infty} v_n e^{-\frac{3}{2}x^2}\, dx = \int_{-\infty}^{+\infty} \{xv_{n-1} - (n-1)\, v_{n-2}\}\, e^{-\frac{3}{2}x^2}\, dx$$

$$= -\tfrac{1}{3}\int_{-\infty}^{+\infty} \frac{d}{dx}(e^{-\frac{3}{2}x^2})\, v_{n-1}\, dx - (n-1)\int_{-\infty}^{+\infty} v_{n-2}\, e^{-\frac{3}{2}x^2}\, dx$$

$$= \frac{n-1}{3}\int_{-\infty}^{+\infty} e^{-\frac{3}{2}x^2} v_{n-2} dx - (n-1)\int_{-\infty}^{+\infty} v_{n-2}\, e^{-\frac{3}{2}x^2}\, dx$$

$$= -\tfrac{2}{3}(n-1)\int_{-\infty}^{+\infty} v_{n-2}\, e^{-\frac{3}{2}x^2}\, dx = -\tfrac{2}{3}(n-1)\,\beta_{n-2}$$

$$= (-\tfrac{2}{3})^{\frac{1}{2}n}(n-1)(n-3)(n-5)\ldots\ldots 1 \times \beta_0,\ n \text{ being of course even.}$$

But $$\beta_0 = \int_{-\infty}^{+\infty} v_0 e^{-\frac{3}{2}x^2}\, dx = \sqrt{\tfrac{2}{3}\pi}.$$

Thus we have the reduction formula:

$$q_{n+2} + \tfrac{1}{2}nq_n = (-\tfrac{2}{3})^{\frac{1}{2}n}(n-1)(n-3)(n-5)\ldots\ldots 1 \times \sqrt{\frac{2\pi}{3}} \quad \ldots\ldots\text{(xxxvi)}.$$

---

* *loc. cit.* p. 5.

† *loc. cit.* p. 4.

We can now rapidly calculate the q's.

$$q_0 = \sqrt{\frac{2\pi}{3}}\frac{\pi}{\sqrt{12}}, \qquad q_2^* = \sqrt{\frac{2\pi}{3}}, \qquad q_4 = -\frac{5}{3}\sqrt{\frac{2\pi}{3}}$$

$$q_6 = \frac{14}{3}\sqrt{\frac{2\pi}{3}}, \qquad q_8 = -\frac{166}{9}\sqrt{\frac{2\pi}{3}}, \qquad q_{10} = \frac{2552}{27}\sqrt{\frac{2\pi}{3}}.$$

This is probably more than sufficient for most practical purposes. Evaluating the coefficients numerically we have from (xxxiii):

$$\frac{p_{22}}{p_{20}^2} = 1 + \cdot 607{,}9271 r^2 + \cdot 140{,}7239 r^4 + \cdot 036{,}7758 r^6$$
$$+ \cdot 010{,}2587 r^8 + \cdot 002{,}9933 r^{10} \quad \text{..........................(xxxvii).}$$

To test the accuracy of this result—obviously correct for $r=0$—consider $r=1$. We have:

$$(p_{22}/p_{20}^2)_{r=1} = 1\cdot 798{,}6788 \quad \text{..........................(xxxviii).}$$

But in the case the variate correlation surface becomes a ridge and $i_1 = i_2$, or:

$$(p_{22})_{r=1} = \frac{1}{N}\int_{-\infty}^{+\infty} i_1^4 \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}} dx$$
$$= \frac{1}{N}\int_{-N/2}^{+N/2} i_1^4\, di_1 = \frac{N^4}{80}.$$

Thus
$$\left(\frac{p_{22}}{p_{20}^2}\right)_{r=1} = \frac{144}{80} = 1\cdot 8 \quad \text{..........................(xxxix).}$$

The difference between (xxxviii) and (xxxix) is only ·001,3212 or about ·07 per cent. Thus even if we omit the term in r^{10}, we shall be less than ·2 per cent in error in this extreme case, when the probable error itself is zero; and for lesser values of r, where the probable error is sensible, we shall not be as much as ·01 per cent in error. This is amply sufficient for statistical purposes. I now take p_{31} and find its value in a different manner.

$$p_{31} = \frac{1}{N}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1^3 i_2 z\, dx\, dy,$$

$$\frac{dp_{31}}{dr} = \frac{1}{N}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1^3 i_2 \frac{dz}{dr}\, dx\, dy = \frac{\sigma_1\sigma_2}{N}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} i_1^3 i_2 \frac{d^2 z}{dx\, dy}\, dx\, dy.$$

This can be integrated twice by parts, and the part between limits vanishes at each integration. Writing $x = \sigma_1 x'$, $y = \sigma_2 y'$ as before, we have:

$$\frac{dp_{31}}{dr} = \frac{3N^4}{(2\pi)^3}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} j_{x'}^2 e^{-\frac{1}{2}(x'^2+y'^2)} z'\, dx'\, dy'.$$

---

* This is found at once from $q_n = 2\int_{-\infty}^{+\infty} j_x e^{-x^2} v_{n-1}\, dx$, or $q_2 = -\int_{-\infty}^{+\infty} j_x d e^{-x^2}$, since $v_{n-1} = x$. Thus

$$q_2 = \int_{-\infty}^{+} e^{-\frac{3}{2}x^3} dx = \sqrt{2\pi}\frac{1}{\sqrt{3}}.$$

The integration with regard to y' can now be completed and we find:

$$\frac{dp_{31}}{dr}=\frac{3N^4}{(\sqrt{2\pi})^5}\frac{1}{\sqrt{2-r^2}}\int_{-\infty}^{+\infty}e^{-\frac{1}{2}\frac{4-r^2}{2-r^2}x^2}j_x^2\,dx \quad\text{..................(xl)},$$

where

$$j_x=\int_0^x e^{-\frac{1}{2}x^2}dx,$$

and we have dropped the dash on x as no longer of service.

Write $m=(4-r^2)/(2-r^2)$, and we must now find:

$$I=\int_{-\infty}^{+\infty}e^{-\frac{1}{2}mx^2}j_x^2dx \quad\text{..............................(xli)}.$$

Now:

$$\frac{dI}{dm}=-\tfrac{1}{2}\int_{-\infty}^{+\infty}x^2e^{-\frac{1}{2}mx^2}j_x^2dx=\frac{1}{2m}\int_{-\infty}^{+\infty}xj_x^2\,d\left(e^{-\frac{1}{2}mx^2}\right)$$

$$=-\frac{1}{2m}\int_{-\infty}^{+\infty}e^{-\frac{1}{2}mx^2}d\left(xj_x^2\right)=-\frac{I}{2m}-\frac{1}{m}\int_{-\infty}^{+\infty}x\,j_x\,e^{-\frac{1}{2}(m+1)x^2}\,dx,$$

$$=-\frac{I}{2m}+\frac{1}{m(m+1)}\int_{-\infty}^{+\infty}j_x d\left(e^{-\frac{1}{2}(m+1)x^2}\right),$$

or:

$$\frac{dI}{dm}+\frac{I}{2m}=-\frac{1}{m(m+1)}\int_{-\infty}^{+\infty}e^{-\frac{1}{2}(m+2)x^2}dx=-\frac{\sqrt{2\pi}}{m(m+1)\sqrt{m+2}},$$

thus:

$$\frac{d}{dm}(\sqrt{m}\,I)=-\frac{\sqrt{2\pi}}{(m+1)\sqrt{(m+1)^2-1}}.$$

Thus:

$$\sqrt{m}\,I=\text{constant}-\sqrt{2\pi}\cos^{-1}\frac{1}{m+1} \quad\text{....................(xlii)}.$$

To evaluate the constant* put $m=1$, and we have:

$$\text{constant}=I_{m=1}+\sqrt{2\pi}\cos^{-1}\tfrac{1}{2}$$

$$=\int_{-\infty}^{+\infty}e^{-\frac{1}{2}x^2}j_x^2\,dx+\sqrt{2\pi}\,\frac{\pi}{3}$$

$$=\int_{-\sqrt{\pi/2}}^{+\sqrt{\pi/2}}j_x^2dj_x+\sqrt{2\pi}\frac{\pi}{3}=\pi^{\frac{3}{2}}/\sqrt{2}.$$

Or finally:

$$I=\frac{\sqrt{2\pi}}{\sqrt{m}}\left\{\frac{\pi}{2}-\cos^{-1}\frac{1}{m+1}\right\}$$

$$=\frac{\sqrt{2\pi}}{\sqrt{m}}\sin^{-1}\frac{1}{m+1} \quad\text{......................................(xliii)}.$$

---

* Mr L. F. Richardson has shown me that if we put $m=0$, since the inverse cosine now vanishes:

constant = Limit $\sigma=\infty$ of $\sqrt{2\pi}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}j_x^2dx$, which he has evaluated with the same result.

Returning now to (xl) we can replace m by its value in terms of r and write

$$\frac{d}{dr}\left(\frac{p_{31}}{p_{20}^{\,2}}\right) = \frac{108}{\pi^2}\frac{1}{\sqrt{4-r^2}}\sin^{-1}\frac{2-r^2}{2(3-r^2)} \quad\ldots\ldots\ldots\ldots\ldots\ldots\text{(xliv)}.$$

This expression I have not succeeded in integrating. I have therefore expanded it in r^2 and then integrated. Since $p_{31}=0$ for $r=0$, we see the constant is zero after integration; thus after some troublesome expansions I find:

$$\frac{p_{31}}{p_{20}^{\,2}} = \frac{54}{\pi^2}\{\cdot 339,8369r - \cdot 005,4820r^3 - \cdot 003,6798r^5 - \cdot 001,1836r^7\}\ldots\ldots\text{(xlv)}.$$

The value of p_{13} is clearly the same as p_{13} for nothing would be altered if x and y were interchanged from (xl) onwards. To test the accuracy of the result, suppose $r=1$. Then we have from the 'ridge':

$$(p_{31})_{r=1} = \frac{1}{N}\int_{-\infty}^{+\infty} i_1^{\,3}\, i_2 \frac{N}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2}x^2/\sigma_1^2}\,dx_1 \text{ and } i_2 = i_1$$

$$= \frac{1}{N}\int_{-N/2}^{+N/2} i_1^{\,4}\,di_1 = \frac{N^4}{80},$$

or:

$$\left(\frac{p_{31}}{p_{20}^{\,2}}\right)_{r=1} = 1\cdot 8, \text{ again.}$$

But (xlv) gives us:

$$\left(\frac{p_{31}}{p_{20}^{\,2}}\right)_{r=1} = 1\cdot 8 \times 1\cdot 00153,$$

that is a result at a maximum only ·15 per cent. in error and correct enough for all statistical purposes. The next step is to determine the powers of r in terms of ρ, and substitute in the expressions just found for $p_{22}/p_{20}^{\,2}$ and $p_{31}/p_{20}^{\,2}$. I find:

$$\frac{p_{22}}{p_{20}^{\,2}} = 1 + \cdot 666,6667\rho^2 + \cdot 108,3084\rho^4 + \cdot 019,7955\rho^6 + \cdot 002,7683\rho^8\ldots\ldots\text{(xlvi)},$$

and:

$$\frac{p_{31}}{p_{20}^{\,2}} = 1\cdot 947,1220\rho - \cdot 123,4135\rho^3 - \cdot 019,4138\rho^5 - \cdot 003,8120\rho^7 \quad\ldots\ldots\text{(xlvii)}.$$

To verify we note that for $\rho=1$, these give 1·7975 and 1·8005 instead of 1·8,—quite sufficiently close for the purpose in view.

We now substitute in (xxxi) and find as far as ρ^8 that:

$$\Sigma_\rho^{\,2} = \frac{1}{N}\left\{1 - 1\cdot 827,5773\rho^2 + \cdot 688,4687\rho^4 + \cdot 112,7773\rho^6 + \cdot 020,2900\rho^8\right\} \text{ (xlviii)}.$$

I throw this, by dividing by $(1-\rho^2)^2$, into the form:

$$\Sigma_\rho^{\,2} = \frac{(1-\rho^2)^2}{N}\left\{1 + \cdot 086,2113\rho^2 + \cdot 012,9408\rho^4 + \cdot 002,3757\rho^6 + \cdot 000,0822\rho^8\right\}^2,$$

or, dropping unnecessary decimals:

$$\Sigma_\rho = \frac{1-\rho^2}{\sqrt{N}}\{1 + \cdot 086\rho^2 + \cdot 013\rho^4 + \cdot 002\rho^6\} \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(xlix)}.$$

Thus we see that the distribution of grades being very far from normal, the probable error $\cdot 67449\Sigma_\rho$ of the correlation of grades exceeds the value $\cdot 67449(1-\rho^2)/\sqrt{N}$, which it would take on the hypothesis of normal correlation by a factor which can amount to about 10 per cent. at a maximum, but gives 0 per cent. excess when $\rho=0$, then agreeing with our previous result.

I propose now to find the probable error in r as determined by grade methods in terms of r. This involves expressing ρ and ρ^2 in terms of r; these are easily found from the known expansions for $\sin^{-1}x$ and $(\sin^{-1}x)^2$. We have:

$$1+\tfrac{1}{2}\rho^2 = 1+\cdot 455{,}9453r^2+\cdot 037{,}9954r^4+\cdot 005{,}0661r^6+\cdot 000{,}8142r^8,$$

$$2\rho = 1\cdot 909{,}8593r+\cdot 079{,}5775r^3+\cdot 009{,}2650r^5+\cdot 001{,}3322r^7.$$

These must be used in (xxxi), which may be written in the form:

$$\Sigma_\rho^{\ 2} = \frac{1}{N}\left\{\left(\frac{p_{22}}{p_{20}^{\ 2}}+1\cdot 8\right)(1+\tfrac{1}{2}\rho^2)-2\rho\,\frac{p_{31}}{p_{20}^{\ 2}}-1\cdot 8\right\}.$$

Hence using (xxxviii) and (xlv), we deduce after some troublesome multiplications:

$$\Sigma_\rho^{\ 2} = \frac{1}{N}\{1-1\cdot 666{,}5507r^2+\cdot 433{,}6130r^4+\cdot 161{,}8337r^6+\cdot 049{,}5042r^8\}$$

$$= \frac{1}{N}(1-r^2)^2\{1+\cdot 333{,}4493r^2+\cdot 100{,}5116r^4+\cdot 029{,}4076r^6+\cdot 007{,}8078r^8\}.$$

But since:

$$r = 2\sin\frac{\pi}{6}\rho,$$

$$\delta r = \frac{\pi}{3}\cos\frac{\pi}{6}\rho\times\delta\rho \quad\text{and}\quad \Sigma_r = \frac{\pi}{3}\sqrt{1-\frac{r^2}{4}}\,\Sigma_\rho.$$

Thus:

$$\Sigma_r^{\ 2} = \frac{\pi^2}{9}\frac{(1-r^2)^2}{N}\{1+\cdot 083{,}4493r^2+\cdot 017{,}1493r^4+\cdot 004{,}2797r^6+\cdot 000{,}4559r^8\}.$$

Taking the square root we have:

$$\Sigma_r = 1\cdot 0472\,\frac{1-r^2}{\sqrt{N}}\{1+\cdot 041{,}7246r^2+\cdot 007{,}7042r^4+\cdot 001{,}8184r^6+\cdot 000{,}1224r^8\},$$

or, for all practical purposes, the probable error of r found from the grade correlation is,

$$\text{P.E. of } r = \cdot 70633\,\frac{1-r^2}{\sqrt{N}}\{1+\cdot 042r^2+\cdot 008r^4+\cdot 002r^6\}\ \ldots\ldots\ldots\ldots\ldots(\text{l}).$$

Clearly for all values of r, this is larger than the probable error of the correlation r found by the product moment method, i.e. $\cdot 67449\,\dfrac{1-r^2}{\sqrt{N}}$. The maximum difference, as r approaches unity is 10 per cent. The value can always be found from (l) without any trouble. The completer value is singularly close to

$$\frac{\cdot 70633}{\sqrt{N}}\,\frac{1-r^2}{\{1-\frac{1}{3}r^2\}^{\frac{1}{8}}}\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(\text{li}),$$

but no advantage is gained in calculation by using this form, as tables of powers of r up to the 6th exist*.

We see therefore from this section that whatever be the value of r, then for normal frequency the probable error of r found by the product moment method is less than the value found by the correlation of grades. Further there is no reason for supposing that the probable error of r found from the difference of grades (R) is not greater than the probable error of r found from the product moment of grades.

We accordingly conclude that the new methods are less accurate than the old. But they possess some advantages,—when ranks can be easily determined,—in rapidity of calculating, and there are undoubtedly cases where they can be used effectively. In saying this I must reassert that I do not believe there is any advantage in the knowledge of rank correlation in itself; I look upon it as a mere stage to the discovery of the variate correlation. For the comparability of rank correlations depends upon the sameness of type in the frequency distributions, and this assumption is the weak step in the method. Granted approximately normal distributions, then the variate correlation flows from the rank correlation, and the whole investigation gains a rich significance.

My remaining sections will be devoted to illustration of the new methods and their comparison with the old.

(9) *Illustration III. Correlation of National Debt and Population.*

The following table is based on data for the year 1900, and raises no pretence to exactness, or financial accuracy. It is merely illustrative.

TABLE IV. *Population and Indebtedness of Various States* 1900.

State	Population in millions	Debt in million £	Population Rank	Debt Rank	$\nu_1-\nu_2$	$(\nu_1-\nu_2)^2$
Russia	129.20	1097·0	1	2	-1	1
United States	76·40	200·0	2	8	-6	36
German Empire†	56·34	649·4	3	4	-1	1
Austria	47·01	226·7	4	7	-3	9
Japan	43·80	51·5	5	15	-10	100
United Kingdom	41·60	705·0	6	3	$+3$	9
France	38·64	1242·1	7	1	$+6$	36
Italy	32·10	500·0	8	5	$+3$	9
Turkey	20·30	162·0	9	9	0	0
Spain	18·10	385·0	10	6	$+4$	16
Belgium	6·82	106·4	11	11	0	0
Roumania	5·50	58·0	12	14	-2	4
Sweden	5·14	18·6	13	17	-4	16
Holland	5·10	95·6	14	12	$+2$	4
Portugal	4·70	155·0	15	10	$+5$	25
Argentine	4·50	86·4	16	13	$+3$	9
Switzerland	3·30	3·6	17	20	-3	9
Greece	2·40	28·0	18	16	$+2$	4
Norway	2·20	12·7	19	18	$+1$	1
Denmark	2·18	11·6	20	19	$+1$	1
			—	—	$30 = S(\nu_1-\nu_2)$	$290 = S(\nu_1-\nu_2)^2$

---

* See *Biometrika*, Vol. II. p. 474.

† Imperial debt and sum of state debts.

Hence: $N^2-1=399$, and $\rho_{12}=1-6\times 290/(20\times 399)=\cdot 7820$.

Further: $R=1-6\times 30/399=\cdot 5489$.

These values are obtained in a few minutes, if the ranks have once been written down. If ρ_{12} only be required, we need not write down the $\nu_1-\nu_2$ column at all, the squares being placed down straight away from the rank columns.

Now applying equations (xxiv) we determine:

$$r=\cdot 7962, \text{ found from } \rho_{12},$$
$$=\cdot 7810, \text{ found from } R.$$

The probable error of r found from ρ_{12} as given by Equation (l) is ·0596. Thus we conclude that

$$r=\cdot 80\pm\cdot 06, \text{ found from } \rho_{12}$$
$$=\cdot 78\pm>\cdot 063, \text{ found from } R^*.$$

If we turn to the much more laborious method of moments, we find:

Mean Population = 27·26 millions; Mean Debt = 289·7 million £,

S. D. Population = 31·74 millions; S. D. Debt = 357·9 million £.

Now these results in themselves should be sufficient to warn us that both distributions are very far from normal; for the S. D.'s in both cases are greater than the means, and since in a normal distribution, we might easily have a deviation equal to the S. D. we should on that hypothesis expect to get negative debts and negative populations. The distributions are therefore very skew, or in clubbing together great and small powers, we have introduced excessive heterogeneity, completely destroying any approach to normality†. If we work out the value of r by the product moment method, we find:

$$r=\cdot 68\pm\cdot 08.$$

We see at once that the rank method has so exaggerated the correlation that it has made the probable error of the less exact methods less than the probable error of the more exact method! The explanation of this lies simply in the fact that the system we are dealing with is not normal. If the ranks of two variables were those given in Table IV, and the distribution were normal, then the variate correlation would be ·80; it actually takes the value ·68, and this is a very good illustration of how much the nature of the distribution may affect a judgment from ranks.

---

* The p. e. is of the form $\frac{\cdot 7738}{\sqrt{n}}(1-r^2)(1+c_1r^2+c_2r^4+c_3r^6)$, the c's being positive unknown constants, and this is $>\cdot 063$.

† If we confine our attention to the seven "great powers," Austria, France, Germany, Great Britain, Italy, Russia and the United States, we find $\rho_{12}=-\cdot 143$, $R=-\cdot 125$, giving $r=-\cdot 15$ and $-\cdot 23$ with a probable error of ·3; this result again emphasises the heterogeneity of the material.

Of course it is doubtful whether when we are in ignorance of the character of the distribution we could say more than

$$r = \cdot 8 \pm \cdot 1, \text{ found from } \rho_{12},$$

and

$$r = \cdot 7 \pm \cdot 1, \text{ found by product-moment.}$$

These might then be treated as identical for some purposes of inference. But the advantage of the longer product-moment method would be that it would have taught us that the correlation was non-Gaussian, and given us in the process the regression line. This would probably more than compensate for its greater laboriousness.

(10) *Illustration IV. Correlation between mean Size of Litter in a Generation and mean Sex Ratio in the same Generation in the case of Mice.*

The following data are taken from a paper in *Biometrika*, Vol. v., p. 439.

TABLE V.

Generation	Mean size of Litter	Mean Sex Ratio	Litter Rank	Sex Ratio Rank	$\nu_1 - \nu_2$	$(\nu_1 - \nu_2)^2$
1st	5·06	·505	5	3	+ 2	4
2nd	4·94	·491	6	4	+ 2	4
3rd	5·96	·523	1	2	− 1	1
4th	5·93	·542	2	1	+ 1	1
5th	5·53	·462	3	6	− 3	9
6th	5·23	·483	4	5	− 1	1

Thus:

$$S(\nu_1 - \nu_2)^2 = 20, \quad S(\nu_1 - \nu_2) = 5,$$

and

$$\rho_{12} = \cdot 429, \quad R = \cdot 143.$$

Whence:

$$r \text{ from } \rho_{12} = \cdot 45 \pm \cdot 23,$$

$$r \text{ from } R = \cdot 25 \pm < \cdot 23.$$

The actual value of r from product-moment is

$$r = \cdot 63 \pm \cdot 17.$$

This example serves to show that the correlation found from R may when the observations are few, not be definitely significant, while when we proceed in the more accurate manner it is definitely significant. The R-method is thus shown not to have special advantages, but rather peculiar disadvantages for short series. Its merit really lies in rapidity of working for assay purposes and rough treatment.

(11) *Illustration V. Resemblance of Cousins.*

(*a*) *Width of Hand.* The following table gives the width of the hand in 34 pairs of male adult cousins taken from my series of Cousin Measurements. These data are being used by Miss Ethel M. Elderton in a forthcoming paper on this subject, and I have most heartily to thank her for the exhaustive manner in which she has dealt with the material in order to illustrate the whole subject of determining correlation by ranks.

Table VI. *Width of Hand in mm. in Pairs of Male Adult Cousins.*

1st cousin x	2nd cousin y	Rank A		Rank B		$\nu_1-\nu_2$		$(\nu_1-\nu_2)^2$	True Grade of A		True Grade of B	
80·7	80·0	23	17	17	23	6		36	23·51	20·74	20·74	23·51
90·0	80·0	58	17	17	58	41		1681	58·82	20·74	20.74	58·82
80·7	84·7	23	48	48	23		−25	625	23·51	40·66	40·66	23·51
90·0	84·7	58	48	48	58	10		100	58·82	40·66	40·66	58·82
80·0	84·7	17	48	48	17		−31	961	20·74	40·66	40·66	20·74
74·5	81·0	3	26	26	3		−23	529	5·52	24·74	24·74	5·52
81·0	80·0	26	17	17	26	9		81	24·74	20·74	20·74	24·74
86·0	81·0	52	26	26	52	26		676	46·00	24·74	24·74	46·00
80·7	83·7	23	43	43	23		−20	400	23·51	36·36	36·36	23·51
94·0	82·7	64	37	37	64	27		729	65·26	31·99	31·99	65·26
94·0	81·7	64	34	34	64	30		900	65·26	27·68	27·68	65·26
76·0	77·0	5	11	11	5		− 6	36	8·44	10·90	10·90	8·44
76·0	79·0	5	16	16	5		−11	121	8·44	17·08	17·08	8·44
76·0	83·0	5	41	41	5		−36	1296	8·44	33·29	33·29	8·44
86·3	88·3	53	54	54	53		− 1	1	47·16	54·16	54·16	47·16
92·5	85·0	60	51	51	60	9		81	63·51	41·94	41·94	63·51
83·7	81·7	43	34	34	43	9		81	36 36	27·68	27·68	36·36
83·7	83·3	43	42	42	43	1		1	36·36	34·62	34·62	36·36
83·7	78·7	43	15	15	43	28		784	36·36	16·05	16·05	36·36
82·0	81·0	36	26	26	36	10		100	28·95	24·74	24·74	28·95
80 5	80·0	22	17	17	22	5		25	22·71	20·74	20·74	22·71
75·0	76·0	4	5	5	4		− 1	1	6·40	8·44	8·44	6·40
71·0	76·0	1	5	5	1		− 4	16	1·70	8·44	8·44	1·70
73·0	77·0	2	11	11	2		− 9	81	3·45	8·44	8·44	3·45
84·5	78·0	47	13	13	47	34		1156	39·81	13·78	13·78	39·81
76·0	78·0	5	13	13	5		− 8	64	8·44	13·78	13·78	8·44
93·3	89·7	61	55	55	61	6		36	64·53	58·09	58·09	64·53
93·3	82·7	61	37	37	61	24		576	64·53	31·99	31·99	64·53
98·7	82·7	66	37	37	66	29		841	67·58	31·99	31·99	67·58
89·7	81·0	55	26	26	55	29		841	58·09	24·74	24·74	58·09
81·0	93·3	26	61	61	26		−35	1225	24·74	64·53	64·53	24·74
82·7	81·0	37	26	26	37	11		121	31·99	24·74	24·74	31·99
98·7	81·0	66	26	26	66	40		1600	67·58	24·74	24·74	67·58
98·7	89·7	66	55	55	66	11		121	67·58	58·09	58·09	67·58
Mean Size } = 83·16		Mean Rank } = 34				$S(\nu_1-\nu_2)$ = 605		$S(\nu_1-\nu_2)^2$ = 2 × 15923	Mean Grade } = 32·41			

The measurements were only read to the millimetre, but since measurements were taken two or three times in each case the fractions ·3, ·5 or ·7 arise, when averaging. Since either cousin may be the "first" cousin, we have for a symmetrical table 68 pairs. In the third and fourth columns, we have the ranks placed, according as to which cousin is considered the "first." It will at once be obvious that many ties arise; thus no less than eight individuals tie with a width of hand 81 mm. at rank 26. It is not so clear what rank ought to be given to them. They run from 26 to 33, we may call them all 29·5. We shall speak of this as the mid-rank method. Or, we might put them all at 26, because this would probably be the result nearest to the true grade*. We shall speak of this as the bracket-rank method†.

The above table illustrates the work for the bracket-rank method in columns 5 and 6, the differences of ranks A and B being, however, only written down *once*, so that to find $S(\nu_1-\nu_2)$, we must sum all quantities in the fifth column as if they had the same sign, and double the sum of their squares in the sixth column.

We find : $R = \cdot 2148$ and $\rho_{12} = \cdot 3922$,

whence r from $\rho_{12} = \cdot 408 \pm \cdot 072$,

r from $R = \cdot 361 \pm > \cdot 072$.

If we now investigate the value of R and ρ_{12} from the mid-ranks, we find that $S(\nu_1-\nu_2) = 588$ and $S(\nu_1-\nu_2)^2 = 29812$. Accordingly:

$R = \cdot 2369$, and $\rho_{12} = \cdot 4310$.

Whence : r from $\rho_{12} = \cdot 448 \pm \cdot 069$,

r from $R = \cdot 396 \pm > \cdot 069$.

Both these values for r are higher than those determined by the bracket-rank process. We must then question whether the mid-rank or the bracket-rank method is the better. Or, indeed is it not possible, that sometimes the one, and sometimes the other will be the closer according to the nature of the frequency distribution?

To illustrate this point the actual grades on the basis of normal distribution have been calculated by Eqn. (xii). It must be remembered that ·5 has to be added to the grade to obtain the rank, Eqn. (xiii).

We find : Mean width of hand = 83·16 mm.

Standard Deviation = 6·201 mm.

As illustration of the method consider the hand of width 84·7 mm., its deviation is 1·54 and the ratio of this to the S.D. = ·248, this corresponds to a value of $\frac{1}{2}(1+\alpha)$, in the notation of Sheppard's Tables, = ·59793 and multiplied by 68 gives the grade 40·66, corresponding to a rank 41·16, as against the observed rank 48 or a mid-rank 49! Thus the actual size of organ corresponding to a bracket rank may differ widely from the size really belonging to the ranked organ, or the true grade in a general population differ very considerably from the spurious grade or rank in the sample used. This point again indicates how little can be judged from ranks unless we associate the rank distribution with some frequency hypothesis.

Having found the true grades we may correlate them together to find ρ_{12}, but in using the formula

$$\rho_{12} = 1 - \frac{S(g_1-g_2)^2}{2N \times \sigma_g^2},$$

---

* That is, find σ_1 and calculate g_1 and g_2 from Eqn. (xii) p. 10; the true grade in this case is 24·74, and $\nu_1 = g_1 + \cdot 5 = 25\cdot 24$ is even below 26, not above it.

† To adopt a term from the examination world, where the place number of the bracket is measured only by those above.

we may adopt either the theoretical value $\frac{1}{12}N^2$ for σ_g^2, or we can actually calculate its value. Now $\frac{1}{12}N^2 = 385\frac{1}{3}$ and $\sigma_g^2 = 365{\cdot}94$, and thus there is a very considerable deviation from normality in the series* $S\ (g_1 - g_2)^2 = 31153{\cdot}195$, and thus:

$$\rho_{12} \text{ found from the true } \sigma_g^2 = {\cdot}3740,$$

$$\rho_{12} \text{ found from } \sigma_g^2 = \tfrac{1}{12}N^2 \quad = {\cdot}4055.$$

Whence :

$$r \text{ from true } \sigma_g^2 = {\cdot}3890,$$

$$r \text{ from } \sigma_g^2 = \tfrac{1}{12} = {\cdot}4215.$$

If we might judge from this single case we should conclude that the bracket-rank method gave a closer result to the grade method than the mid-rank method. But the question now arises, how close after all are all these grade rank methods to the correlation coefficient in any short series such as the present?

Accordingly the series was worked out by product moment and the result obtained was

$$r = {\cdot}331 \pm {\cdot}073.$$

Thus we see that the actual correlation is considerably lower than that given by any of the rank or grade processes. It is perfectly true that ·33 and ·45 are within double the probable error, and therefore two different random samples of the real population might have given as widely divergent results. But this is really the case of two different methods applied to the *same* sample. And further the actual correlation tells us that as far as this sample is concerned the true answer is likely to lie between ·19 and ·48, but the mid-rank method tells us that it is likely to lie between ·31 and ·58†. Now it is clear we might for some extraneous reason hold the value likely to be ·56, and we should find nothing to contradict this in the mid-rank result. But the proper method of determining r would show us that such a value was itself very unlikely. Thus the latter method when it diverges less than twice the probable error from the result of the rank method may yet forbid us to interpret the results in a manner admissible on the rank method. We cannot argue in like manner from the grade or rank result because that method has assumed an hypothesis, not made in the product-moment treatment, i.e. that of normal correlation, which is here not justified by the results.

But even the amount of agreement here noted is to be considered rather exceptional. I owe to Miss Elderton the working out of three other pairs of characters in the same set of male cousins each in five different ways. I have myself done each of them in three more ways, namely by Variate Differences as in Art. 2, and by the R method. The results are given in the Table below.

---

* The mean grade in fact = 32·41 and not 34 also.

† Taking a range of twice the probable error on either side the means.

TABLE VII. *Comparison of Correlation Coefficients found by Various Methods. Resemblance of Hand in* 68 *Pairs of Male Cousins.*

Character	Product Moment	Variate Difference	Grades		Ranks			
					Bracket-Rank		Mid-Rank	
			True σ_g^2	$\sigma_g^2=\frac{1}{12}N^2$	By ρ_{12}	By R	By ρ_{12}	By R
Width of Hand	·33 ± ·07	·37	·39	·42	·41 ± ·07	·36	·45 ± ·07	·40
Width of Wrist	·17 ± ·08	·25	·12	·22	·07 ± ·085	·05	·08 ± ·085	·03
Length of Index Finger	·19 ± ·08	·14	·16	·13	·21 ± ·08	·29	·19 ± ·085	·29
Length of Little Finger	·29 ± ·075	·26	·18	·30	·20 ± ·08	·19	·24 ± ·08	·21
Mean of Four Results	·25	·25	·21	·25	·22	·22	·24	·23
Root Mean Square Deviation from true r		·053	·069	·060	·079	·094	·079	·096

It will, I think, be clear from this table that for series even with as many as 68 pairs—and this is approaching the limit at which any time is gained by using rank methods—we cannot hope to ascertain the correlation of the *sample* by such methods within about ·1 of its value, and as the probable error of the sample may be ·07, we may well deviate ·2 from the population value in our estimate. We are accordingly very unlikely to reach reliable results by rank methods for the 8 to 10 observations to which Dr Spearman proposes to apply his R-method. We see that the mean values are fairly close, although the variate difference and the second grade methods give the best results. Judged by mean square deviations from product moment results, the variate difference is easily first, then come the laborious grade methods, the rank methods by ρ_{12} about fifty per cent. worse than the variate difference, and lastly the R methods not quite 100 per cent. worse. Thus we note that when a series is not fairly long and not approximately normal, the different rank and grade methods will give very diverse results. But when a series is fairly long, say 100 or more observations, then there is no advantage in rapidity from the rank method; the formation of a grouped correlation table, and the use of the product moment is just as rapid, and further conveys a great deal more of valuable information.

(12) *Conclusions.* Three new methods of determining variate correlation have been given in this paper. The first, that of variate differences, seems likely to be of some service in the case of symmetrical tables containing large numbers, the frequency being approximately normal, homotyposis tables may be taken as illustration.

The second that of deducing variate correlation from correlation of ranks, may be of service when it is not possible to put a quantitative value on the individual character. Thus it might be easy to form a relative series of intensity of pigment, and place individuals in rank. But mere correlation of ranks is not in itself a comparable character, as the variate correlation may have widely different values for the same ranking. Justification for the comparability depends upon assuming a wide spread rule of frequency distribution, and this rule can hardly be other than normality. The present paper shows how to deduce variate correlation from correlation of ranks. It shows, however, that such a method of reaching variate correlation is considerably less exact than the usual product-moment method. There is no gain in accuracy, but the reverse in using such a method in the case of short series.

Thirdly, the method proposed by Spearman of deducing the correlation of ranks from the positive differences of ranks is discussed, and the error of the process by which he has deduced for it an accuracy greater than that of the more usual methods of finding correlation is indicated. A method for deducing variate correlation from positive difference of ranks is indicated. The method is very rapid for short series, say those not exceeding 20 observations, but it is less accurate than the product-moment method, and considerable changes in the final value reached will be found to arise according as we use bracket-ranks or mid-ranks in the case of ties. The comparison with true grades for a few special cases, does not enable us to say which is the better method; the deviations from normality sometimes appear to make one, sometimes the other, the closer to the true correlation.

In conclusion, I think, we may say that variate correlation found by ranks may prove to be a useful *auxiliary* method of dealing with correlation, when it is needful to give a rough answer to a problem in a brief time, or when the material itself is incapable of being accurately measured. In all such cases mean square of rank differences will be more accurate than mean positive rank difference. But both methods must be used with caution, and their easy application must not lead us to approve exaggerated statements as to their accuracy.

CAMBRIDGE: PRINTED BY JOHN CLAY, M.A. AT THE UNIVERSITY PRESS.

Published by Dulau and Co., Soho Square, London, W.

DRAPERS' COMPANY RESEARCH MEMOIRS.

DEPARTMENT OF APPLIED MATHEMATICS, UNIVERSITY COLLEGE, UNIVERSITY OF LONDON.

These memoirs will be issued at short intervals.

Technical Series.

I. On a Theory of the Stresses in Crane and Coupling Hooks with Experimental Comparison with Existing Theory. By E. S. Andrews, B.Sc.Eng., assisted by Karl Pearson, F.R.S. *Issued.* Price 3s.

II. On some Disregarded Points in the Stability of Masonry Dams. By L. W. Atcherley, assisted by Karl Pearson, F.R.S. *Issued.* Price 3s. 6d.

III. On the Graphics of Masonry Arches, with special reference to the Relative Strength of Two-pivoted, Three-pivoted and Built-in Metal Arches. By L. W. Atcherley and Karl Pearson, F.R.S. *Issued.* Price 6s.

IV. On Torsional Vibrations in Axles and Shafting. By Karl Pearson, F.R.S. *Issued.* Price 6s.

V. An Experimental Study of the Stresses in Masonry Dams. By Karl Pearson, F.R.S., and A. F. Campbell Pollard, assisted by C. W. Wheen, B.Sc.Eng., and L. F. Richardson, B.A. *Issued.* Price 7s.

VI. On the Graphics of Masonry Structures. By N. G. Dunbar, Chadwick Scholar, and Karl Pearson, F.R.S. *Shortly.*

Biometric Series.

I. Mathematical Contributions to the Theory of Evolution.—XIII. On the Theory of Contingency and its Relation to Association and Normal Correlation. By Karl Pearson, F.R.S. *Issued.* Price 4s.

II. Mathematical Contributions to the Theory of Evolution.—XIV. On the Theory of Skew Correlation and Non-linear Regression. By Karl Pearson, F.R.S. *Issued.* Price 5s.

III. Mathematical Contributions to the Theory of Evolution.—XV. On the Mathematical Theory of Random Migration. By Karl Pearson, F.R.S., with the assistance of John Blakeman, M.Sc. *Issued.* Price 5s.

IV. Mathematical Contributions to the Theory of Evolution.—XVI. On Further Methods of Determining Correlation. By Karl Pearson, F.R.S. *Issued.* Price 4s.

V. Mathematical Contributions to the Theory of Evolution.—XVII. On Homotyposis in the Animal Kingdom. By E. Warren, D.Sc., A. Lee, D.Sc., Edna Lea-Smith, Marion Radford and Karl Pearson, F.R.S. *Shortly.*

Studies in National Deterioration.

I. On the Relation of Fertility in Man to Social Status, and on the changes in this Relation that have taken place in the last 50 years. By David Heron, M.A. *Issued.* Price 3s.

II. A First Study of the Statistics of Pulmonary Tuberculosis. By Karl Pearson, F.R.S. *Issued.* Price 3s.

PUBLICATIONS OF THE FRANCIS GALTON LABORATORY FOR NATIONAL EUGENICS, UNIVERSITY OF LONDON. (Published by Dulau and Co.)

I. The Inheritance of Ability. Being a Statistical Examination of the Oxford Class Lists from the year 1800 onwards, and of the School Lists of Harrow and Charterhouse. By Edgar Schuster, M.A., First Francis Galton Research Fellow in National Eugenics, and E. M. Elderton, Galton Research Scholar in National Eugenics. *Issued.* Price 4s.

II. A First Study of the Statistics of Insanity and the Inheritance of the Insane Diathesis. By David Heron, M.A., Second Galton Research Fellow. *At Press.*

III. The Promise of Youth and the Performance of Manhood, being a Statistical Examination into the Relation existing between Success in the Examinations for the B.A. Degree at Oxford and subsequent Success in Professional Life. (The professions considered are the Bar and the Church.) By Edgar Schuster, M.A., First Galton Research Fellow in National Eugenics. *Issued.* Price 2s. 6d.

IV. On the Degree of Resemblance of First Cousins. By Ethel M. Elderton, Galton Research Scholar. *At Press.*

Published by the Cambridge University Press.

BIOMETRIKA.

A JOURNAL FOR THE STATISTICAL STUDY OF BIOLOGICAL PROBLEMS.

Founded by W. F. R. Weldon, Francis Galton and Karl Pearson.

Edited by Karl Pearson in Consultation with Francis Galton.

Vol. V., Part III.

I. A Biometrical Study of Conjugation in Paramecium. (With eleven Diagrams in the text.) By Raymond Pearl, Ph.D.

II. The Anthropometric Characteristics of the Inmates of Asylums in Scotland. (With eleven Diagrams and eight Maps in the text and three Plates of Maps.) By J. F. Tocher.

III. On the Error of Counting with a Haemacytometer. By Student. (With two Diagrams in the text.)

Miscellanea: (i) On the Distribution of Severity of Attack in Cases of Smallpox. By F. M. Turner, M.D. (ii) Remarks on Dr Turner's Note. By Karl Pearson, F.R.S.

Supplement to Vol. V. Anthropometric Survey of the Inmates of Asylums in Scotland. (With Maps.) By J. F. Tocher. Issued by permission of the Henderson Trustees.

Vol. V., Part IV.

I. Statistical Observations on Wasps and Bees. By F. Y. Edgeworth.

II. Natural Selection in *Helix Arbustorum*. By A. P. di Cesnola.

III. Grades and Deviates. By Francis Galton, with a Table of Deviates by W. F. Sheppard.

IV. A Coöperative Study of Queens, Drones and Workers in *Vespa Vulgaris*. By A. Wright, A. Lee and Karl Pearson, F.R.S.

V. Statistical Studies in Immunity. A Discussion of the Means of estimating the severity of cases of Acute Disease. By John Brownlee, M.D., D.Sc.

VI. On Heredity in Mice from the Records of the late W. F. R. Weldon. Part I. On the Inheritance of the Sex-Ratio and of the Size of Litter.

VII. The Calculation of the Moments of a Frequency-Distribution. By W. F. Sheppard.

Miscellanea: (i) On the Inheritance of Psychical Characters. Being further Statistical Treatment of Material Collected and Analysed by Messrs Heymans and Wiersma. By Edgar Schuster, M.A., and Ethel M. Elderton. (ii) Reply to certain criticisms of Mr G. U. Yule. (With coloured Plate XXIII.) By Karl Pearson, F.R.S.

Notices and Bibliography.

The subscription price, payable in advance, is 30*s.* *net* per volume (post free); single numbers 10*s.* *net.* Volumes I., II., III., IV. and V. (1902–7) complete, 30*s.* *net* per volume. Bound in Buckram 34*s.* 6*d.* *net* per volume. Subscriptions may be sent to C. F. Clay, Manager, Cambridge University Press Warehouse, Fetter Lane, London, E.C., either direct or through any bookseller.